Plath, Ingrid

Understanding meta-analyses. A consumer's guide to aims, problems, evaluation and developments

Baden-Baden : Nomos 1992, 136 S. - (Studien zum Umgang mit Wissen; 7)

Contents

Preface		7
1	**Introduction**	8
1.1	Traditional reviewing	10
2	**What is meta-analysis?**	11
2.1	Aims and functions of meta-analysis	12
2.2	The value of the meta-analytic approach	14
3	**Criticized aspects of meta-analysis**	16
3.1	Objectivity	16
3.2	Sampling bias	17
3.2.1	Publication bias	17
3.2.2	Selection bias	19
3.2.3	The quality of primary studies	20
3.3	Quantitative and statistical aspects	22
3.3.1	Mixing apples and oranges	22
3.3.2	Effect sizes: non-independence and other aspects	24
3.3.3	Critique of techniques and focus of the statistical analysis	25
4	**Issues of reliability and validity**	29
4.1	Reliability	29
4.2	Validity	31
4.2.1	Internal validity	32
4.2.2	Statistical conclusion validity	33
4.2.3	Construct validity	34
4.2.4	External validity	34
5	**Guidelines for evaluating meta-analyses**	37
5.1	Recommendations on how to evaluate a meta-analysis	40
5.2	Construct and external validity	41
5.2.1	Problem formulation and hypothesis selection	41
5.2.2	Sampling and selection of studies	42
5.2.3	Presenting the characteristics of the primary studies	43
5.2.4	Interpreting results	44

5.3	Internal and statistical conclusion validity	45
5.3.1	Coding	45
5.3.2	Statistical analysis	46
5.4	Of what use is the evaluation?	48
6	**Evaluation of a sample of meta-analyses**	49
6.1	Method	49
6.1.1	Sample description	49
6.1.2	Coding procedure	50
6.1.3	Data evaluation	51
6.2	Results	51
6.2.1	Theoretical framework	55
6.2.2	Sampling	56
6.2.3	Coded study characteristics	58
6.2.4	Data analysis	60
6.2.5	Interpretation	64
6.3	Do critiques and replication attempts affect evaluations?	66
6.4	Is the quality of meta-analyses improving?	67
6.4.1	Problematic issues: sample size and representativeness	67
6.4.2	Problematic issues: coding procedure	69
6.4.3	Possible trends in the reporting quality of meta-analyses	70
6.5	Resumé	75
7	**Improving the quality and utility of meta-analyses**	77
7.1	Combining qualitative and quantitative reviewing approaches	78
7.2	Taking communication quality into account	79
7.3	Knowledge synthesis and practical relevance	80
7.4	Reviews of reviews -- meta-synthesis?	85
7.5	Concluding remarks	87
Glossary		91
References		98
Appendix		109

1 Introduction

Practitioners have several paths open to them when wanting to find practical, scientifically based solutions to problems encountered during the course of their work or wishing to gain insight into the present state of research in a particular field of interest. They can ask experts for advice. They can search databases for relevant empirical studies. They can consult published reviews or books. All these approaches have their pros and cons, especially concerning their feasibility, representativeness, objectivity, reliability and validity.

Traditional literature reviews have long served as a relatively convenient source of information. Since 1976 a specific form of review, called among other things quantitative synthesis, quantitative research integration or meta-analysis, has enjoyed increasing popularity with scientific reviewers. KULIK (1984) estimates that after eight years about a 1000 papers have been published on the topic, about a third of which are actual reviews. Since then the number has been steadily increasing.

Propagated as an objective, scientific way of research integration, newcomers optimistically approach this type of review in the hope of finding well founded answers to their questions. Accustomed to traditional reviews the novice is likely to be slightly overwhelmed. Confronted with masses of quantitative data, numerous statistical analyses, controversial discussions of statistical issues involved and relatively limited substantive information, the review will possibly cause more confusion than transmit actual information.

Newcomers can react to this state of affairs in several ways. They can stick to traditional reviews, despite the weaknesses they might have, prefering the more easily intelligible narrative report. They can stick to meta-analyses, skipping the complicated parts to read the conclusions, trusting the expert's evaluations without being able to reconstruct how these were reached. They can also decide to delve into the methodology of quantitative syntheses, becoming semi-experts themselves and thus able to grasp the finer points of the review and its conclusions.

The last alternative is clearly what meta-analysts expect of potential consumers. To quote BANGERT-DROWNS (1986, p.388): "Readers, researchers, editors and research reviewers need to be better informed if they are to be intelligent consumers and critics of meta-analytic reviews." This sounds simpler than it is. The relevant methodological literature is usually directed at readers interested in conducting their own meta-analyses. Information specifically useful to consumers of reviews is gleaned mainly as by-product from the numerous methodological papers spread over diverse journals and a few introductory books.

Instead of becoming simpler the subject tends to become more complicated and complex as one progresses more deeply into it. One is confronted with methodological arguments, controversies, criticisms, praises and doubts. What was hailed as the approach to clarify and systematically analyze the diversity of empirical research findings in particular domains, is now ironically adding to the confusion.

Under these circumstances the question arises whether a quantitative review still has relevance for a practitioner. Does one really, as WALBERG (1984) suggests, have to gain insight into the explicit, quantitative, empirical techniques used in the current mainstream of theory and research if one wishes to use or understand it? Can one expect practitioners additionally to take this task upon themselves or is it the expert's responsibility to make scientific knowledge available to users in an acceptable form? Why should the consumer undertake the arduous task of studying the technical details of the approach if the simpler alternative of traditional reviews exists? No guidelines were ever required to read these and having done so one usually had the impression of having gained some insight into the theory and findings of the reviewed domain. This is not always the case with meta-analyses. Rather, one feels that to understand and profit from these one should already have been well versed in the methodology and the theory of the subject covered by the integration.

What previous knowledge does the consumer really need to be able to work effectively with meta-analytic reviews and to critically evaluate their quality? It is contended, that highly detailed knowledge of the technical and statistical issues is not absolutely necessary for this purpose, rather a clear conception of the aims or intentions of this type of reviewing and problems encountered, limiting or preventing their attainment. This knowledge is not only useful for working with quantitative syntheses but also encourages a more sensitive handling of traditional reviews.

Having in part experienced the quandary described above and deciding to resolve it by working through the relevant sources to penetrate the details of the approach, the aims of the present report are: to focus on the needs of the potential consumer, to summarize scattered methodological information considered essential, to present recommendations for the critical evaluation of the approach, to discuss its practical relevance and to describe the impressions gained studying actual meta-analyses on specific educational topics. Hopefully, this will help reduce the apparently widespread lack of familiarity with the approach which appears to be a matter of general concern (WANOUS, SULLIVAN & MALINAK, 1989). Since consumers rather than potential conductors of meta-analyses are the primary focus, technical and statistical details have intentionally been kept at a minimum, yet the use of and reference to technical concepts and procedures is unavoidable. For this reason terms possibly unfamiliar to persons not well-versed in statistics and empirical research methodology have been briefly outlined in the glossary.

1.1 Traditional reviewing

Research integration and conflict resolution are essential steps in the process of knowledge accumulation and refinement, linking past research with future scientific endeavours (PILLEMER & LIGHT, 1980). Reviews have long served this purpose: to summarize accumulated information concerning a specific topic, highlighting the resolved and unresolved issues by generalizations drawn from a set of studies pertaining directly to the area of interest (COOPER, 1982; JACKSON, 1980).

Ideally, the review should replace papers fallen behind the research front and direct future research (COOPER, 1982). Unfortunately the traditional, qualitative or narrative way of reviewing led to conflicting results which in turn led to increased disenchantment with and criticism of this type of reviewing. Repeatedly criticized shortcomings are subjectivity or bias, e.g. neglecting large amounts of information and extracting it inefficiently (COOK & LEVITON, 1980; COOPER & ROSENTHAL, 1980; LIGHT & PILLEMER, 1982). This concerns firstly, the imprecise weighting of conclusions with regard to the proportion of research it is based on or covers and secondly, the overemphasis of significant results, usually disregarding the actual magnitude and direction of effects. Similar criticism is leveled at one of the first attempts to improve reviewing practices, the so-called box or vote counts: making frequency counts of the number of favourable, neutral or adverse statistical significance tests contained in the studies reviewed to obtain a more objective impression of the general trend of results. While also ignoring the actual effect sizes, a reviewer trying to make sense of such data is actually engaged in, as MEEHL (1978, p.823) scathingly puts it, "meaningless substantive construction on properties of the statistical power function, and almost nothing else." In general, critics seem to agree with the conclusion reached by COOPER and ROSENTHAL (1980) that some of the confusion and contradiction is due to the manner of integration and not necessarily a function of the results.

Part of the problem is the lack of explicit methods and rules to guide the task of reviewing a body of empirical studies (FISKE, 1983; JACKSON, 1980). Another is the vast number of studies potentially available to a reviewer, making the integration of results almost impossible and difficult without the application of adequate statistical techniques (also cf. GLASS, 1977).

As a response to this unsatisfactory state of affairs (STRUBE & HARTMANN, 1982) as well as the growing realization that research synthesis is a scientific enterprise just as important as primary research or theory development (PILLEMER & LIGHT, 1980), the approach to research integration referred to as meta-analysis was developed.

2 What is meta-analysis?

First introduced in 1976 by Glass, meta-analysis is characterized as involving "the attitude of data analysis applied to quantitative summaries of individual studies" (GLASS, McGAW & SMITH, 1981, p.21) or "the application of research methods to the characteristics and findings of research studies" (p.23, ib.). "It is not a technique; rather it is a perspective that uses many techniques of measurement and statistical analysis" (p.21, ib.). Or to quote STRUBE and HARTMANN (1983, p.14): "Meta-analysis is not simply a collection of quantitative techniques. Rather, it represents a systematic approach to the problem of integrating a common research domain."

Fundamental to this research approach to accumulating knowledge is that reviewing requires the application of rigorous scientific standards to and formal procedures for combining the results of empirical research evidence (BULLOCK & SVYANTEK, 1985; PILLEMER & LIGHT, 1980). This also implies that one ensure the potential intersubjective testability or replicability of the reported results (GLASS, McGAW & SMITH, 1981). Meta-analysis seeks generalizable answers and regular patterns from the divergent findings of numerous individual research studies on a given topic by statistical analysis (BANGERT-DROWNS, 1986; KULIK & KULIK, 1989). Besides this, the investigation of how findings vary from one study to the next and the influence of methodology constitutes a substantial part of the meta-analysis (GLASS & KLIEGL, 1983; GLASS, McGAW & SMITH, 1981).

The main features distinguishing this approach from the traditional form of reviewing are:

- Finding a large, if possible, comprehensive sample of studies pertaining to the substantive issue (KULIK & KULIK, 1988, 1989)
- Transforming the study results to a common metric, focusing on the effect size rather than just the statistical significance of results (CORDRAY & ORWIN, 1983; KULIK & KULIK, 1989)
- Describing the study features in quantitative or quasi-quantitative terms by utilizing a coding scheme (KULIK & KULIK, 1988, 1989).
- Applying statistical techniques to these summary statistics: aggregating the findings across studies as well as systematically examining the relations between study features and outcome (CORDRAY & ORWIN, 1983; KULIK & KULIK, 1988, 1989; LEVITON & COOK, 1981).
- Systematic and detailed description of the whole process of integration to ensure replicability (JACKSON, 1980).

In keeping with the view of research integration as a strictly scientific endeavour is the conceptualization of meta-analysis in analogy to primary research as involving a series of stages: problem formulation, hypothesis selection, sampling of studies, definition and measurement of variables, data analysis and interpretation, reporting or publication of results (COOPER, 1982; FRICKE & TREINIES, 1985; GLASS, McGAW & SMITH, 1981; JACKSON, 1980; WANOUS, SULLIVAN & MALINAK, 1989). Furthermore, to keep in this spirit of inquiry, one should not read a meta-analysis without establishing some opinion of its quality. Each of the steps is open to criticism, especially as they involve a series of decisions on the part of the reviewer. Just as in primary research, using similar criteria, one will have to analyze its reliability and validity.

Most recently DRINKMANN (1990) attempted to formalize a definition of meta-analysis which combines all the elements mentioned so far: emphasizing both its statistical, quantitative characteristics and its structural similarity to primary empirical research and criteria (p.11: "Metaanalyse soll sein: eine an den Kriterien empirischer Forschung orientierte Methode zur quantitativen Integration der Ergebnisse empirischer Untersuchungen sowie zur Analyse der Variabilität dieser Ergebnisse.") Although the above descriptions contain references to the aims and functions of meta-analysis, these will now be examined in greater detail.

2.1 Aims and functions of meta-analysis

The aims of meta-analysis do not necessarily differ from those of traditional qualitative reviews. The approach makes the difference: the application of statistical tools to reach conclusions and the strict adherence to scientific standards in the sense of ensuring potential intersubjective testability.

The focus is on summarizing research evidence and rendering it useful to the consumers of research: the lay public, policy makers, practitioners like teachers as well as scientists and undergraduates (COOPER & ROSENTHAL, 1980; PILLEMER & LIGHT, 1980). As much information as possible should be drawn from existing data in a systematic manner (PILLEMER & LIGHT, 1980), establishing consensus, identifying outliers and moderators of findings (GREEN & HALL, 1984). Special attention is given to conceptual and technical issues as the conclusions usually are to be generalized to a broader field (PILLEMER & LIGHT, 1980). The emphasis is on practical simplicity (GLASS, McGAW & SMITH, 1981), necessitating the formulation of uncomplicated conclusions from the statistically based integration, else these would be lost on all but a few specialists (LIGHT & PILLEMER, 1982).

Focusing on these general aims a carefully conducted meta-analysis can fulfil a number of functions (STRUBE & HARTMANN, 1982). The summary of the functions

given below largely follows the distinctions and definitions given by STRUBE and HARTMANN (1983). They represent the ideal hoped to be achieved.

Self-evident and fundamental is the *descriptive function*. The meta-analysis should give a clear and detailed summary of what we know, how we got to know it and the quality of this knowledge by special consideration of the methodological, procedural and theoretical variables involved. Fulfilling the remaining functions depends almost entirely on how carefully and comprehensively this purpose was accomplished.

Meta-analysis can serve as *guide for planning future research*. Presenting an accumulation of research methods and procedures formerly used either with or without success in the research domain as well as highlighting the characteristics moderating the effects, can help scientists to plan their research more effectively. (At the same time they should, however, remember that if no methodological variables were found to mediate the results, this does not mean that any methodology can be applied successfully.)

The *diagnostic function* of meta-analysis is closely allied to the previous one. It is diagnostic in the sense that it identifies gaps in the knowledge, highlights well investigated areas and uncovers flaws. To cite STRUBE and HARTMANN (1983, p.23): "Meta-analysis can help to identify 'holes' in the 'nomological net' of multiple empirical relationships that constitute a theory."

The *function of transmitting information to practitioners* is just as important as presenting it to fellow scientists. The reviewer is seen in the role of a gatekeeper. Deciding whether the accumulated research evidence is ready for practical application and the domains to which it can be applied, depends on the detailed investigation of existing studies with special attention directed toward the aspects of external validity.

The *predictive function* of meta-analysis reaches beyond the actual findings of the individual studies. According to STRUBE and HARTMANN (1983) this potential use of reviews is often neglected. Because of its statistical approach, meta-analysis is well suited to this purpose. Looking upon each reviewed study as an independent datapoint, allows the examination of plausible hypotheses not tested in the single studies (COOK & LEVITON, 1980; PILLEMER & LIGHT, 1980). Classifying the studies along some dimension of interest, the relation of this newly constructed variable to study outcomes can be determined, as for instance when searching for possible explanations of contradictory findings or moderators of results. More clearly predictive use of studies as datapoints is made when employing regression analysis to estimate the effect of hypothetical values of the independent variables.

Looking toward the future STRUBE and HARTMANN (1983) even visualize a *generative function* of meta-analysis. The extensive, systematic summarization potentially

allows the creative restructuring of the information thereby gaining new theoretical insights, i.e. generating new theories.

2.2 *The value of the meta-analytic approach*

Closely related and often identical to the above are the frequently reported strengths, advantages, contributions or benefits of meta-analysis (cf. COOK & LEVITON, 1980; FISKE, 1983; GREEN & HALL, 1984; HEDGES, 1986; JACKSON, 1980; KAVALE, 1988; KULIK & KULIK, 1988; PILLEMER & LIGHT, 1980).

Meta-analysis forces the reviewer to take an active approach to literature, focusing attention on how to find, organize, classify and draw conclusions from research evidence. Besides highlighting the poor methodological quality of previous reviewing practices, meta-analysis led to the realization that even though all steps undertaken in the process of research integration involve qualitative judgements, they can be completely specified, thus likely to be more objective and allowing replication.

Furthermore, the development of adequate statistical tools paved the way for efficient and extensive evaluations of the vast number of potentially available studies in certain research domains. Converting the study findings into effect sizes having a common metric allows the estimation of overall effects as well as the systematic examination of the variation of effect sizes across various study characteristics. The application of statistical techniques enables one to make constructive use of contradictory findings, e.g. examining the influence of variable definitions, research design, analysis strategy, unit of analysis or measurement technique as possible mediators of results. The combination of data-sets, having similar construct definitions of relevant variables, increases the power of statistical tests as well as allowing a more accurate description of the relations between variables, if these have been implemented at different levels.

Indirectly, meta-analysis influences the quality of primary studies by showing up poor reporting practices and inadequate conceptualization and theorizing. Ultimately, the value of meta-analysis rests on whether it helps understanding a complex body of empirical findings (GUSKIN, 1984). KULIK and KULIK (1989) see the ultimate test of the worth of a research methodology in the contribution it makes to our understanding: accumulating knowledge, moving research and development forward, producing results that can serve as guide to action. In this connection David Hilbert is frequently cited as having said that the importance of a scientific work is measured by the number of previous publications it makes superfluous to read (GLASS, McGAW & SMITH, 1981; FRICKE & TREINIES, 1985; HORNKE, 1983).

Meta-analysis can help reduce the confusion present in the steadily increasing, heterogeneous research literature, if used by those knowledgeable about the substance and methodology in the field of integration (GUSKIN, 1984). KULIK and KULIK (1989) express a similar point of view, stating more definitely that meta-analysis transforms findings from something confusing and contradictory into something that can be useful to a variety of audiences, hoping that scientists, administrators or teachers use the conclusions to guide future research, policy formulation and action. Like COOPER and ROSENTHAL (1980), KULIK and KULIK (1989) believe that conclusions supported by a carefully conducted meta-analysis have a precision, clarity and force they lack when expressed as an opinion in a narrative review. The conclusions will appear to be more rigorous and objective.

There seems to be a general consensus that the major contribution of meta-analysis is that it is objectively verifiable, using measured concepts, quantitative data and statistical analysis (BULLOCK & SVYANTEK, 1985). The concrete formulation of rigorous scientific ideals and specific intentions keyed to both scientists and practitioners invites the reader of meta-analyses to take a critical stance, to examine whether the actual meta-analysis lives up to its ambitious aims and functions. Understandably, the problems encountered with meta-analysis and the criticisms directed at the approach concern exactly these aims and functions as well as its supposed benefits or advantages.

3 Criticized aspects of meta-analysis

The striking feature of methodological discussions of meta-analysis, that its potential strengths are seldom mentioned without also pointing out its potential dangers, is an outgrowth of the critical attitude induced. Yet, the general tenor of these caveats is that the limitations or problems are not intrinsic to the approach, but arise due to thoughtless, uncritical or inefficient practices (ABRAMI, COHEN & d'APOLLONIA, 1988; COOPER & ARKIN, 1981; LEVITON & COOK, 1981; STRUBE & HARTMANN, 1982). A meta-analysis does not have to be flawed if sufficient attention is given to possible pitfalls and one is sensitive to the assumptions and often complex decisions involved (COOK & LEVITON, 1980; KULIK & KULIK, 1989; LEVITON & COOK, 1981; SLAVIN, 1984; STRUBE & HARTMANN, 1982). Consequently, the critical points are usually presented together with their potential solutions or advice on how to avoid or counteract them. A critical consumer of meta-analysis should be aware of these because they can affect all stages of the meta-analytic process and are inseparable from the formulated aims.

In addition to unique statistically based problems, the approach is recognized as encompassing all those also encountered in traditional reviewing (STRUBE & HARTMANN, 1983; STRUBE, GARDNER & HARTMANN, 1985). The main objections concern the claim to objectivity, the possibility of bias and various critical aspects of the quantitative and statistical nature of meta-analysis. A summary of the arguments repeatedly voiced follows.

3.1 Objectivity

Meta-analysts, having criticized the subjectivity of traditional reviews, are not surprisingly confronted with the same objection. The meta-analytic process is replete with judgements and decisions at all levels of abstraction: deciding on what hypotheses to test; which studies to include or exclude; how to deal with missing information; which effect sizes to calculate and how; what study features to code and how etc. (COOK & LEVITON, 1980; HAGER, 1984; MINTZ, 1983). Instead of objective it would be more accurate to say that the approach requires judgemental decisions of an equally subjective kind, but as these are made explicit by decision rules (MINTZ, 1983), they can be challenged, if deemed necessary, or replicated to test their adequacy (KULIK & KULIK, 1989; WANOUS, SULLIVAN & MALINAK, 1989). One can determine the reliability of the judgements and consequently estimate to what extent the conclusions reached can be trusted.

Critics seem to fear that the apparent objectivity and allure of the quantitative approach will blind readers to the potential limitations or lend a 'patina of science' to conclusions which could be based on unreliable or invalid data and procedures. This potential scientism is, however, not inherent in the approach, but in how it is used (COOPER & ARKIN, 1981; SLAVIN, 1984; WORTMAN, 1983). Properly conducted, all decisions and judgements involved are explicitly stated to ensure intersubjective testability. Inadequate reporting, so often criticized by meta-analysts in primary research reports, is, however, just as prevalent in meta-analytic reports, a fact which makes replications difficult and leads to seemingly different findings (WANOUS, SULLIVAN & MALINAK, 1989). Misuse cannot be criticized as limitation of a research methodology. The user is to blame. Meta-analysis has sensitized reviewers to the complex decisions involved in reviewing. This should help to prevent that meta-analysis is seen as some sort of infallible objective procedure that can just be pulled off the shelf to plug data in (LEVITON & COOK, 1981) and alert the reader to beware if these judgemental aspects are not described in detail.

The problems of objectivity and bias are closely linked and should be considered simultaneously. For the sake of clarity, however, the two are discussed separately.

3.2 Sampling bias

Just as in the previous case, bias, one of the main objections to traditional reviews, is now seen as a potential danger by critics of the meta-analytic approach. Subsumed under the term sampling bias are three related yet distinguishable forms of bias: publication bias, selection bias and quality of the primary studies. The latter is not usually discussed under the heading of bias, but as the presentation will show, can be considered as a special form of selection bias.

Sampling is a crucial step in the meta-analytic process because its adequacy determines the range of valid generalizations, but a good sampling plan may be thwarted by such matters as publication or selection bias (HEDGES, 1986).

3.2.1 Publication bias

The question of publication bias is critical because meta-analyses are largely based on published data (BULLOCK & SVYANTEK, 1985; KULIK & KULIK, 1988). The meta-analyst depends on the public availability of data and research processes. This is one of the reasons why the majority of studies included in meta-analyses are published articles. Less often data are obtained from dissertations, technical reports or convention presentations, even more rarely from file drawer studies relevant to the topic (STRUBE

& HARTMANN, 1982). The retrieved studies might not reflect the population of studies actually conducted (ROSENTHAL, 1984). In this sense the sample may be unrepresentative, nonrandom and biased. A question often investigated in this respect is whether effect sizes calculated from published articles differ from those obtained from unpublished ones. As the results do not always coincide (FRASER, WALBERG, WELCH & HATTIE, 1987; GLASS, McGAW & SMITH, 1981; KULIK & KULIK, 1989; ROSENTHAL, 1984), this possible source of between study effect size variation should be examined in meta-analyses.

The problem of relying primarily on published information can also be looked at from a slightly different angle. Because of certain publication policies these studies may have an overabundance of significant results, or contain most Type I errors, i.e. false positive results (KRAEMER & ANDREWS, 1982; STRUBE & HARTMANN, 1982). The extreme view taken is that the published studies represent the 5% false significant results whereas the 95% nonsignificant results remain locked up in some drawer.

Aptly dubbed the file drawer problem, ROSENTHAL (1979) developed a method (the fail-safe-test) to estimate the extent of this problem. By calculating the number of nonsignificant studies necessary to raise the overall probability of a Type I error to 0,05 an index is obtained, allowing one to estimate whether publication bias due to Type I errors is a plausible rival hypothesis for a significant combined probability level. ROSENTHAL (1979) was able to show that whereas for large samples of studies the file drawer problem does not seem to be a critical problem, this is not the case for small samples. Only a few nonsignificant studies filed away suffice to make the combined significant p of the reviewed studies nonsignificant. In the meantime, however, experts have pointed to various disadvantages of the technique and raised doubts as to its statistical adequacy (SCHÖNEMANN, 1990).

A related problem is the question of bias in the reporting of the primary study itself. Selective reporting of statistical analyses and their results or incomplete descriptions of the treatments, controls, experimental procedures and outcome measures can invalidate the integration. These issues are referred to more generally as the problem of missing data. Random missing data would not present a substantial problem but, as HEDGES (1986) points out, there is little reason to believe that these are random. Rather, they could be systematically related to effect sizes or important characteristics of the study, thus a potential threat to the validity of meta-analytic conclusions.

The problem of publication bias is restricted to the stage of the meta-analytic review concerned with finding research literature and formulating a strategy to obtain a comprehensive pool of studies from which those eventually included in the meta-analysis are selected. The adequacy with which this is done will, however, affect the quality of the whole meta-analysis.

3.2.2 *Selection bias*

One might ask, why select at all after having found a comprehensive pool of studies, thus avoiding the reproach of selection bias. The meta-analytic approach is a quantitative one, requiring the coding of study features and calculation of effect sizes. For both purposes detailed information is needed, stressing once more the influence missing data could have on the quality of the meta-analysis.

Some primary studies contain insufficient data for the calculations (JACKSON, 1980). Others contain reporting inaccuracies, making the coding difficult if not impossible (STRUBE & HARTMANN, 1982). Furthermore, the more specific and detailed the coding procedure is, the greater the problem of missing data will become (HEDGES, 1986). Some of the retrieved studies will be theoretically irrelevant to the questions posed by the meta-analyst. Others will be weaker or stronger tests of the theory (COOK & LEVITON, 1980). These problems necessitate the formulation of inclusion and exclusion criteria to guide the selection process without unnecessary bias and guarantying intersubjective testability. Any criticism directed at the study selection process can then be demonstrated empirically by reanalysis (FISKE, 1983; for a practical example cf. BRYANT & WORTMAN, 1984).

At present the meta-analytic methodology seems to be applicable mainly to comparative studies, i.e. studies comparing the treatment effects obtained in different experimental groups, and investigations of the correlation between two variables (CORDRAY & ORWIN, 1983). As yet it appears less capable of accommodating research based on single-subject designs, time-series analyses, qualitative or case studies and those involving complex statistical designs (DRINKMANN, 1990) although procedures are being developed (e.g. for single-subject research, cf. SCRUGGS, MASTROPIERI & CASTO, 1987). This will inevitably lead to the selection of a particular section of research evidence. To allow an assessment of the representativeness of the sample of studies, authors should include a list of the selected studies and also report those excluded along with the reasons for their rejection as SLAVIN (1986) suggests. Furthermore, DRINKMANN (1990) was able to show that deciding on a specific meta-analytic methodology can possibly lead to systematic selection effects: when using simple techniques the database is larger than when complex meta-analytic methods are employed. Directional bias (e.g. retaining more studies supporting the hypotheses) and bias towards more highly significant results could play an important role in this reduction. For this reason he advocates a multi-method approach which would allow these effects to be studied in detail if present.

Moreover, ignoring studies that do not fit the conceptual or methodological framework developed for coding could result in the elimination of important innovative studies. The coding scheme should accommodate the largest portion of relevant lit-

erature, dealing qualitatively with those that do not fit in or contain insufficient data for calculating effect sizes (GREEN & HALL, 1984; also cf. STANLEY, 1987). However, this tendency to maximize the number of studies could lead to the inclusion of some only minimally relevant or inappropriate to the subject being integrated (GUSKIN, 1984).

As SLAVIN (1984) points out, uncritical inclusion criteria may remove the personal selection bias of the reviewer, but not the bias in the underlying studies. This objection introduces one of the most controversial aspects of meta-analyses: the problem of the quality of primary studies selected for the synthesis.

3.2.3 The quality of primary studies

The currently widespread practice of including all studies meeting certain pre-established criteria regardless of quality in the meta-analysis is the main objection (KULIK & KULIK, 1988; SLAVIN, 1984). Critics point out that a synthesis is only as strong as the data that go into it. No statistical technique can extract valid and reliable conclusions from data of poor quality (HEDGES & OLKIN, 1985; SLAVIN, 1984). For this reason some suggest that only 'good' studies be included in the meta-analysis.

Not selecting studies on the basis of quality is justified by pointing out that there are no simple and definite rules for deciding, what constitutes a good method or measure (HUNTER, SCHMIDT & JACKSON, 1982; JACKSON, 1980; WORTMAN, 1983). The inclusion or exclusion criteria could thus be biased by personal predilections (KRAEMER & ANDREWS, 1982). Advocates of unrestricted inclusion hope to circumvent problems by coding various aspects of study quality and then analyzing the covariance of these features with study results. Whether design or quality make a difference is explicitly tested, the notion being that if flaws are crucial they will be correlated with study findings (GLASS & KLIEGL, 1983), assuming furthermore that the studies can converge on 'true' conclusions if they are imperfect in different respects, i.e. contain no systematic bias (COOK & LEVITON, 1980; GLASS, McGAW & SMITH, 1981).

This solution has not been considered as acceptable by all, the reproach being that meta-analysts tend to underplay the methodological problems common to a body of research (GUSKIN, 1984). Furthermore, as HEDGES (1986) points out, not all threats to validity are equally important in every specific research area, recommending attention to the salient methodological difficulties and the use of multiple criteria to establish the quality of the studies. Meta-analyses cannot of themselves detect bias in predominantly one direction (COOK & LEVITON, 1980; JACKSON, 1980). Neither coding nor statistical analysis take care of this problem. So even if results across studies seem similar,

they should be critically analyzed to detect possible methodological flaws common to all (GUSKIN, 1984; HEDGES, 1984; JACKSON, 1980; SLAVIN, 1984).

Apart from this the analysis of covariation may prove difficult because the number of studies per coded category is not large enough (COOK & LEVITON, 1980; JACKSON, 1980). Furthermore, studies excluded from the meta-analysis for reasons of insufficient statistical information are frequently also the ones methodologically worst, causing a restriction of the range of quality covered, consequently making the results of the statistical analysis questionable (WORTMAN, 1983). As BRYANT and WORTMAN (1984) suggest, the decision on whether to use the all-inclusive or selective approach should depend on the range of methodological quality present in the particular research domain being reviewed.

MINTZ (1983) points to another difficulty. The meta-analyst is usually interested in the relations between various variables. Therefore the conclusion that quality is not related to effect size should rest on the fact that it does not affect the relationship between these variables, and not by establishing that quality is nonsignificantly related to one of them, as is predominantly done.

Another aspect of the problem of quality does not immediately concern the primary studies. Coding itself may affect the quality of the data. Valid and reliable coding is difficult to achieve (JACKSON, 1980). Errors slip in when coding and the problem of inaccurate reporting is unresolved (GREEN & HALL, 1984; STRUBE & HARTMANN, 1982, 1983). Additionally, the quality of coding is influenced by reporting style and thoroughness. These differences can be confounded with actual differences between study findings, making the interpretation of the analysis difficult (BULLOCK & SVYANTEK, 1985). Furthermore, the categories for coding aspects of quality are often fairly broad and selective, thus precluding an exact and detailed analysis of their influence (WORTMAN, 1983).

As before, critics fear that the apparent specificity of results lend an unwarranted sense of security and precision to the findings of meta-analyses. For this reason the potential sources of bias should be discussed and analyzed in detail, stating decision rules explicitly, to ensure replicability, objectivity and enable the critical examination of statistical techniques used to determine the effects of quality and design.

The problems presented so far are not unique to meta-analysis but also confront the traditional reviewer to a large extent. The meta-analyst, however, goes on to quantify and statistically analyze the sample of studies selected. The methods involved are the target of what constitutes the third major category of criticism directed at the approach. Some of these have already been mentioned above, as they are applied to help solve the

problems encountered during these stages of the meta-analysis. Others repeatedly, criticized, will be summarized in the following section.

3.3 Quantitative and statistical aspects

The objections range from a general condemnation of the quantitative approach as such, to highly specialized criticism of the specific statistical theory and formula involved, sometimes culminating in suggesting alternative synthesis approaches such as SLAVIN's (1986) best evidence synthesis or HAGER's (1985) approach emphasizing the evaluation of the validity of substantive hypotheses rather than concentrating on the testing of statistical hypotheses. The following discussion will concentrate on those aspects readily comprehensible by non-experts, e.g. what has become known as the apples and oranges problem, critical points concerning the effect sizes and other statistical problems of a more general kind.

3.3.1 Mixing apples and oranges

Just as controversial and as widely discussed as the issue of quality, the apples and oranges problem concerns the practice of meta-analysts to aggregate effect sizes across studies. Transforming study results to a common metric allows one to combine measures covering entirely different things (KULIK & KULIK, 1988). Critics argue that averaging across diverse operationalizations of treatment (independent) and outcome (dependent) variables results in combined effect sizes whose meaning is distilled, uninterpretable and difficult to fathom (KULIK, 1984; SLAVIN, 1984). For example, the independent variable could be cooperative learning. What exactly cooperative learning entails in the individual studies might differ completely. The same applies to a dependent variable such as achievement which could for example be measured as grade at the end of the year, as performance in a class test or on some standardized achievement test.

The summation across diverse outcomes is considered the more serious part of the problem (KULIK, 1984). The conclusions drawn from these averages are usually presented in the form of statements of the following kind: the treatment groups show x,xx standard deviation superiority as compared to the control groups. But the pertinent question is, what do they actually show superiority of (KULIK, 1984; KULIK & KULIK, 1988). SLAVIN (1984) goes so far as to say that the conclusions reached add up to nothing more than the claim that a treatment designated as experimental groups has more positive effects than a treatment designated as control groups.

Related to the above is the criticism of meta-analyses for using very broad definitions and tending to respond to questions no longer asked (COOK & LEVITON, 1980; MINTZ, 1983). Broad construct definitions entail the danger of disregarding theoretical

relevance while broad questions additionally lack practical relevance, both increasing the likelihood of confounded, uninterpretable results.

GLASS, McGAW and SMITH (1981) respond to these arguments by pointing out that only studies varying to a certain degree create the need for integration. Good generalizations are obtained by ignoring distinctions that make no important difference (GLASS, 1983). The sensibility of averaging depends on the question asked, just as the distinctions made or ignored depend on the choice of object field and taxonomy, i.e. what is to be explained or understood and how the variables are defined (GLASS & KLIEGL, 1983). So, it might for instance be sensible to ignore the differences between various types of therapies if the question is whether some form of therapy is better than none.

A practicable way to obviate these objections would be to use broad constructs, but to distinguish narrower constructs in the data analysis or presentation of results (HEDGES, 1986). The general way out proffered for all problems of heterogeneity whether of constructs, measures, study characteristics or quality, is to analyze subgroups comparable on theoretical or conceptual grounds separately or by blocking, excluding or weighting them according to clear criteria (KULIK, 1984; ROSENTHAL, 1984). SLAVIN (1984) maintains, however, that heterogeneity is more often ignored than explored as recommended. Pooled, single, inadequately explained and oversimplified effect sizes are emphasized with a tendency to gloss over details and a resulting loss of information (ROSENTHAL, 1984).

In the same vein critics comment on meta-analysts' overemphasis of statistical details rather than focusing on critical substantive issues specific to the field being integrated (HAGER, 1984; SLAVIN, 1984).

In general, critics seem to feel that meta-analysis could make a more useful contribution if more attention were given to the methodology and substance of studies as well as to the questions or hypotheses it is intended to answer rather than tending towards blind empiricism (GUSKIN, 1984; HORNKE, 1983; STUBE & HARTMANN, 1982). The quantitative nature of meta-analysis encourages a tendency to consider methodological and substantive problems as dealt with once they have been coded, which can lead to distortion and misinterpretation (GUSKIN, 1984; STRUBE & HARTMANN, 1982).

Warnings that meta-analysis cannot replace thought, planning and intelligent discussion of crucial issues repeatedly accompany these criticisms together with the indication that meta-analytic findings can only enhance and aid the interpretation, not determine it (SLAVIN, 1984; STRUBE & HARTMANN, 1982).

Problems created by aggregating effect sizes across studies can be solved more or less satisfactorily. This is not the case for other critical aspects encountered with effect sizes.

3.3.2 *Effect sizes: non-independence and other critical aspects*

One of the more persistent problems pertaining to effect sizes concerns the issue of non-independence or so-called 'inflated N'. Individual empirical studies frequently test several hypotheses on the same sample of subjects. More than one effect size can thus be calculated per study. These are not independent and artificially inflate the sample size (N) if they, instead of studies, are used as unit of analysis (GLASS, 1977; GLASS, McGAW & SMITH, 1981; GREEN & HALL, 1984; HEDGES, 1986; KULIK & KULIK, 1988; ROSENTHAL, 1984; STRUBE & HARTMANN, 1982). Another form of non-independence can arise when individual studies were conducted by the same researcher or originate from the same laboratory or team. Ideosyncratic methodology, atypical and regional samples of subjects or experimenter bias may lead to a sample of studies not truely independent (HEDGES, 1986; ROSENTHAL, 1984; STRUBE & HARTMANN, 1982).

Both types of non-independence create persistent and as yet unsolved difficulties in the statistical analysis. Estimating the error contained in the statistics describing the data-set is problematic, determining the true correlation among study features nearly impossible. As the data fail to meet assumptions for valid significance testing and regression analysis these have to be interpreted with extreme caution (GLASS, McGAW & SMITH, 1981; GREEN & HALL, 1984; HEDGES 1986; KULIK, 1984; KULIK & KULIK, 1988).

Another widely discussed problem of effect sizes concerns the formula used to calculate them. The difficulty stems from the fact that primary studies lack a common metric and use diverse designs (JACKSON, 1980). The formula are criticized as not giving unbiased estimates (HEDGES, 1980) or for ignoring various critical factors (for detailed information refer to the lucid exposition by KULIK & KULIK, 1986).

Especially deplored is the widespread tendency to overlook the need for a variety of formula to calculate effect sizes and standard errors and neglecting the critical distinction between interpretable and operative effect sizes introduced by Cohen (KULIK & KULIK, 1986, 1988, 1989; also cf. COOPER, 1981; GLASS, McGAW & SMITH, 1981). While interpretable effect sizes calculated from differently designed studies are conceptually equivalent and interpreted on a common scale, operative effect sizes are not (the standardizing units differ). This causes artifactual relations between effect sizes and experimental design. On top of this, variations in the choice of metric result in substantially different effect sizes (McGAW & GLASS, 1980). Therefore, to achieve un-

flawed analyses and conclusions, interpretable effect sizes using formula which take study design, sample size and test statistic into account and are expressed in the common metric of final status scores have to be calculated. As McGAW and GLASS (1980) point out, using final status as scale has three advantages: 1) it is readily interpretable and phenomenologically relevant; 2) it is more reliable than using derived measures and 3) findings expressed in other scales can be converted to final status measures but not vice versa.

In general, to be interpretable the effect size must be meaningful to the statistically unsophisticated, be transportable and comparable across different designs and measures (KRAEMER & ANDREWS, 1982). Effect sizes are descriptions of the degree to which a relation departs from the null state (COOPER, 1981). However, ROSENTHAL (1984) doubts that the practical importance of effect sizes is clear even to statisticians. To make it intuitively meaningful he developed the binomial effect size display which shows the percentual changes in success rate due to treatment (cf. ROSENTHAL, 1982; ROSENTHAL & RUBIN, 1982b). To assist the interpretation of the effect sizes Cohen's quantitative classification scheme for small, medium and large values is sometimes called upon (COOPER, 1981; WORTMAN, 1983). The general wisdom of this practice is considered doubtful (COOPER, 1981; GLASS, McGAW & SMITH, 1981; WORTMAN, 1983). To obtain a meaningful interpretation of an effect size one should rather compare it to the magnitudes of effect sizes previously obtained in the research domain (normative approach) or use experts to assess the practical significance of the value (judgemental approach) (COOPER, 1981; SECHREST & YEATON, 1981; WORTMAN, 1983; also cf. KENDALL & MARUYAMA, 1985).

The discussion of the methodological problems encountered in meta-analyses tends to become increasingly technical. Interested readers should consult the original papers as the main aim of the present one is not to develop guidelines for potential meta-analysts, but to alert the consumers of meta-analyses to difficulties encountered and identified.

3.3.3 Critique of techniques and focus of the statistical analyses

The actual statistical techniques applied in meta-analyses vary and experts do not agree as to which are the most appropriate. The standard package includes estimating a combined probability, average effect sizes, stability of results and factors associated with differential treatment outcomes (STRUBE & HARTMANN, 1982). For each of these purposes a variety of procedures is available. The problem facing both conductors and consumers of meta-analyses is that to date there is little firm basis for evaluating the relative utility and appropriateness of these alternatives (DRINKMANN, 1990; STRUBE & MILLER, 1986).

Conventional statistical procedures such as ANOVA, t-test or regression analysis are considered inappropriate because meta-analytic data do not meet the criteria necessary for their valid application (GLASS, McGAW & SMITH, 1981; GUSKIN, 1984; HEDGES, 1984, 1986; HEDGES & OLKIN, 1985; STRUBE & HARTMANN, 1982). Alternative approaches developed in part to circumvent these problems (e.g. HEDGES & OLKIN, 1985; HUNTER, SCHMIDT & JACKSON, 1982; ROSENTHAL, 1984; SLAVIN, 1986) are in turn criticized, in particular by KULIK and KULIK (1986, 1988, 1989), among other things for advocating inappropriate statistical methods and formula for testing the influence of study features on outcomes and calculating effect sizes.

BANGERT-DROWNS (1986) identifies five forms of meta-analytic methods currently in use which he differentiates according to differences in purpose, unit of analysis, treatment of study variations and outcomes of the analysis. These five approaches also reflect the gradual evolution and refinement of methodology triggered by criticism of the former techniques.

However, categorizations tend to obscure the fact that these strategies are not necessarily mutually exclusive (PILLEMER & LIGHT, 1980). They can be used in parallel to investigate whether similar conclusions are obtained from each, which seems a sensible thing to do as long as the methodological issues associated with each are not solved. In fact, DRINKMANN (1990), after systematically comparing meta-analytic findings obtained when analyzing the same database by different statistical methods, concludes that taking a multi-method approach in meta-analysis is advisable (also cf. STRUBE, GARDNER & HARTMANN, 1985).

Another aspect frequently criticized is the focus on main effects and the assessment of only relatively direct evidence on a topic. Interaction effects also reported in the primary studies are largely ignored, the reason being that few of these involve the same or comparable factors, e.g. while one study might investigate the differential influence of prior knowledge or motivation on experimental effects, others might concentrate on the influence of gender, socio-economic status or type of school subject (COOPER & ARKIN, 1981; GREEN & HALL, 1984). Ignoring interaction effects is problematic, because then the reader does not even know they exist. They should at least be reported. The theoretical variables involved can also be coded as study feature (GREEN & HALL, 1984). Disaggregating studies according to these at least reflects a sensitivity to the problem (COOPER & ARKIN, 1981) and reduces the likelihood of concentrating on the examination of nominal, demographic or gross empirical characteristics when searching for possible moderators of effect sizes (COOK & LEVITON, 1980). In response to the as yet unsatisfactory handling of interaction effects in meta-analyses DRINKMANN (1990) developed more or less practicable alternative techniques to approach the problem, which readers should refer to if interested.

Conflicting findings result from differences in treatment, setting and participant characteristics as well as design and analysis features of primary studies (JACKSON, 1980; LIGHT & PILLEMER, 1980). Additionally, the use of different formula for calculating effect sizes can produce confounding variation. Statistical techniques applied in meta-analyses can neither identify specific features as causes of between study variations in effect size nor can they untangle inherently correlated (confounded) events (JACKSON, 1980; HEDGES, 1986; MINTZ, 1983). They can only help indicate which aspects could sensibly be examined experimentally as possible mediators of results. In this connection the danger of over-interpreting or over-analyzing the results is pointed out. To avoid the problem, the heterogeneity of the data should be established before searching for mediating variables, i.e. the observed sample variance should be greater than the variance that can be expected solely due to chance variations (FRICKE & TREINIES, 1985; GREEN & HALL, 1984).

A problem frequently overlooked is the capitalization on chance (KULIK & KULIK, 1989). Meta-analysis usually involves the coding and analysis of dozens of variables. Therefore chance alone will provide some statistically significant relations (BULLOCK & SVYANTEK, 1985), or as HEDGES (1986) puts it, there is a definite limit to the number of tests that can be supported by a given collection of studies. The analyses should thus all be reported to enable an estimate of the potential gravity of the issue. One possible way of coping with the problem is to develop the hypotheses to be tested before coding and to restrict coding to the relevant variables (BULLOCK & SVYANTEK, 1985; HEDGES, 1986). A related aspect one should keep in mind is that as the number of studies increases so does the likelihood of credible significant conclusions based on spurious findings (BULLOCK & SVYANTEK, 1985; ROSENTHAL, 1984).

Conclusions reached by meta-analyses are misleading and imprecise if statistical techniques or formula are misapplied (HEDGES & OLKIN, 1985; STRUBE, GARDNER & HARTMANN, 1985; STRUBE & HARTMANN, 1983; KULIK & KULIK, 1989). But even if this were not the case, the question of their generalizability has as yet not been solved satisfactorily. It is not clear whether the studies reviewed can be looked upon as a representative, random, unbiased sample or whether they should be considered as the universe of studies (JACKSON, 1980; MINTZ, 1983; STRUBE & HARTMANN, 1983).

Furthermore, the unit of analysis is usually either the study or the study findings. It is not clear whether this allows the generalization of conclusions to people (MINTZ, 1983). According to LIGHT and PILLEMER (1984), by including all studies available on a subject one can generalize to the population of study outcomes. To generalize to a larger population of individuals the studies reviewed have to use representative samples of participants. However, whether one will be able to generalize at all depends on the va-

lidity and reliability of all steps involved in the research integration as well as the clarity of the definitions of the constructs concerned.

To some extent the arguments voiced for or against the meta-analytic approach resemble and remind one of the wellknown protracted controversy between proponents of qualitative and quantitative methodology. A resolution of the conflict is not in sight, but a definite weakening of the respective fronts (cf. CAMPBELL, 1987; FROMM, 1990; LAMNEK, 1988; LINDBLOM, 1987; REICHARDT & COOK, 1979; RESTIVO & LOUGHLIN, 1987; WORTMAN, 1983b). The advocates on either side are beginning to see that the limitations of the one may be the benefits of the other and that the antithesis could be used profitably.

A similar insight is prevalent in reviewing circles. The quantitative and qualitative approaches complement one another (cf. CHELIMSKY & MORRA, 1984; COOK & LEVITON, 1980; COOPER & ROSENTHAL, 1980; HEDGES, 1986; LIGHT & PILLEMER, 1982; SLAVIN, 1984, 1986). Using both will improve the process of research integration. On the one side there is the realization that some things cannot be quantified without considerable loss of practical relevance and information, on the other that techniques applied and conclusions reached can benefit from the consideration of the concepts of reliability and validity.

Both concepts are regarded as essential in evaluating quantitative, empirical research. Not surprisingly, the focus is similar when determining the quality of meta-analyses. For this reason some of the issues involved will be summarized in brief.

4 Issues of reliability and validity

All the objections to and criticisms of meta-analysis mentioned so far can in fact be regarded as problems concerning its reliability and validity. The two concepts just provide a different more theoretical framework or perspective for essentially the same difficulties.

In the present context the issue of reliability and validity can be approached on two levels: that of the primary study and that of the meta-analysis itself, the latter being the primary focus of the present discussion. The two levels are not independent. As indicated by the heated debate on the inclusion or exclusion of 'bad' quality studies, reliability and validity are considered crucial factors affecting the quality of meta-analyses. The concepts of reliability and validity cannot strictly be separated either, reliability being a necessary but not sufficient condition for the validity of a study. Nonetheless, the two will be discussed independently as far as possible.

To recapitulate: meta-analysis was conceptualized as involving a series of stages, each requiring specific decisions: problem formulation; definition of constructs; sampling, selection and coding of studies; data analysis and interpretation; and publication. The question of reliability and validity can be addressed to each of these.

4.1 Reliability

Reliability refers to the consistency, stability and replicability of procedures and measurements, i.e. will the same or similar results be obtained when the same procedures and measurements are repeatedly employed. Evidently one condition for reliability can be fulfilled by ensuring that meta-analytic reports contain enough information on strategies and decision rules to allow replication (cf. section 3.1 on objectivity). The issue most often discussed under the heading of reliability, however, is the accuracy of measurement.

Measurement in meta-analysis is the quantification and coding of study characteristics (GLASS, McGAW & SMITH, 1981). Consequently, much thought has gone into how to improve this process. Suggestions range from advising critical discussions with colleagues about the variables to be coded (ROSENTHAL, 1984) to developing extensive code books, training coders, conducting pilot-studies on coder-source reliability to improve code books if necessary, retraining coders and rechecking reliability (MATT, 1989; STOCK, OKUN, HARING, MILLER, KINNEY & CEURVORST, 1982). The rating of study quality needs special attention, its reliability being difficult to achieve, es-

pecially if overall quality judgements are required rather than rating specific aspects (FISKE, 1983; STOCK, OKUN et al., 1982).

As ORWIN and CORDRAY (1985) point out, however, no amount of training or improving coding rules can eliminate disagreements due to deficient microlevel reporting, i.e. the clarity, completeness and adequacy of the individual investigator's report. Moreover, the authors were able to show that the reliability varied considerably across coding items, thus casting doubt on the usual practice of reporting average or global indices of intercoder agreement, reliability or consistency. Furthermore, the confidence raters had in their judgements varied in a corresponding way. Thus, uncertainty leads to unreliable coding. Omission of information in primary studies leads to the specification of coding conventions which might consistently over- or underestimate true values. In addition, meta-analysts would be hard put to correct and diagnose recording errors in primary studies which typically occur at a rate of 1% but can on occasion reach 48% (ROSENTHAL, 1984), quite apart from the fact that they themselves are also bound to make some inadvertent calculation and editing mistakes (cf. CORDRAY & ORWIN (1983) and HEDGES (1986); for possible ways to counteract these deficiencies; BRYANT & WORTMAN (1984) discuss possible problems of bias encountered in the process).

Although investigations of the reliability of meta-analyses seem to concentrate on the coding stage, some other aspects have also been studied. MATT (1989) examined the reliability of four decision rules frequently employed for selecting primary study effect sizes: conceptual redundancy, coder agreement, outcome reliability and outlier truncation rules. The necessity to select arises from the fact that studies often have several treatment and control groups, measured at different points in time, resulting in multiple redundant effect sizes (compare 'inflated N', p.24). Apart from not being specific enough to allow accurate replication, the decision rules lead to different average effect size values, thus once more stressing the need for explicit and detailed publication of rules to guide the complex decisions coders have to make.

Even though it is difficult to determine the reliability of locating studies or study retrieval (ROSENTHAL, 1984), factors influencing the accuracy of judgements concerning a study's relevance for a meta-analysis have been examined (COOPER & RIBBLE, 1989). Although the results were not conclusive, the accuracy of relevance judgements appears to profit from previous experience in searching for literature and publishing reviews. High conceptual complexity and tolerance of ambiguity on the part of the searcher also seems to enhance the process as well as having the abstract rather than just the title as source of information.

All in all, the problems of reliability appear to be far from solved and more complex than evident on first sight. Unreliable data can jeopardize the validity of the whole meta-

analysis. Therefore the least one should expect of meta-analytic reports is that they provide detailed information on all procedures and rules employed, discuss problems encountered in coding and with primary study reporting practices, and present reliability coefficients for the individual items coded rather than just average indices.

4.2 Validity

The concept of validity refers to the question of the true meaning of the variables coded, measured or analyzed (GLASS, McGAW & SMITH, 1981), i.e. are those things theoretically intended really being tested or investigated. As coding provides the basic data for this process it should present an accurate picture of the literature on the research topic (KULIK & KULIK, 1989). Truth can, however, only be approximated, never definitely established (COOK & CAMPBELL, 1979).

It is not the intention here to enter deeply into the complex theoretical discussion on validity types and their interrelationships nor to expostulate on which 'labels' to use for practically identical issues of validity (for details cf. BRINBERG & McGARTH, 1982; COOK & CAMPBELL, 1979; FRICKE & TREINIES, 1985; JUDD & KENNY, 1982; WORTMAN, 1983b). The terminology employed by COOK and CAMPBELL (1979) seems adequate to examine the two questions of primary interest when exploring the validity of studies: Were the outcomes really caused by the interventions? For which persons, settings, treatment and outcome variables do the effects hold? The first refers to issues of internal and statistical conclusion validity, while the second concerns construct and external validity.

Even more than in the case of reliability, the validity of a meta-analysis depends on the quality of the individual studies in this respect. The two levels are inextricably entangled. On the one hand study features potentially threatening or enhancing the validity constitute a large portion of the items coded, thus directly influencing the validity of meta-analyses which use these as independent, blocking or stratifying variables. On the other hand the validity types are employed as selection and exclusion criteria of studies (cf. SLAVIN, 1986; STRUBE & HARTMANN, 1982, 1983). So, for example, WORTMAN and BRYANT (1985) suggest using the external and construct validity of a study to determine its relevance for the meta-analysis and then the internal and statistical conclusion validity of the relevant study to determine its acceptability for the analysis (cf. BRYANT & WORTMAN (1984) for an empirical examination of the consequences of using this exclusion/inclusion procedure).

Except for the first position, this order of selection has similarities with the relative priorities among types of validity proposed by COOK and CAMPBELL (1979) for applied research: internal validity, external validity, construct validity of effects, statistical

conclusion validity and construct validity of causes. The validity types are interrelated. Trying to increase one will probably decrease another (also cf. BRYANT & WORTMAN, 1984). In order to minimize trade-offs their relative importance should be determined subject to the specific purpose of the study. To accomplish this, one will have to know what the various validity types refer to.

4.2.1 Internal validity

Internal validity concerns the actual research process, the quality of the design used, the correspondence between theory and its implementation. The question of causal attribution is of primary importance. Was, for example, a study conducted in such a way as to allow the conclusion that the experimental intervention caused the outcome and not some other competing or confounding factors?

In meta-analyses the internal validity of the reviewed studies has been the main focus of discussions. All important aspects of the internal validity of primary studies should be coded so that their possible influence on results can be analyzed. Most important among these are the adequacy of assignment rules or selection, the control techniques or groups employed and instrumentation problems. Any threat to the internal validity of primary studies can potentially invalidate meta-analyses, e.g. if an independent variable correlates with an assignment variable confounding is bound to occur. Any systematic bias will cause problems for the meta-analysis (cf. section 3.2 on bias). The detailed evaluation of the internal validity of the reviewed studies is thus a prerequisite for ensuring the internal validity of the meta-analysis.

Decisions made in the course of the reviewing process itself also affect the internal validity of the integration. Foremost among these are the issues discussed under the heading of reliability which can be looked upon as problems of instrumentation and selection. What features are coded, what effect sizes included, which of several types of control groups possibly used in the primary studies is chosen as reference and what criteria employed for study selection are some aspects bound to influence the adequacy with which meta-analytic hypotheses are tested.

Closely linked to the problem of reliability and internal validity are the matters discussed under the label of statistical conclusion validity. Both concern the design of studies, but whereas internal validity focuses primarily on the question of systematic bias, statistical conclusion validity explores the question of random errors (COOK & CAMPBELL, 1979).

4.2.2 *Statistical conclusion validity*

The issue investigated under the heading of statistical conclusion validity is whether appropriate statistical techniques have been used for the analysis of data. Were the assumptions met? Does the method parallel the question asked? Is it powerful enough to detect relationships between independent and dependent variables, should these exist?

Unless primary studies are adequate in this respect, a meta-analysis cannot be expected to deliver valid conclusions. But equally important are the techniques employed in the meta-analysis itself. Judging from the amount of literature published, this seems to be the validity type meta-analysts have been concerned with predominantly (cf. section 3.3 on the criticism of the quantitative and statistical aspects of meta-analyses).

Numerous articles have been published on the appropriateness of meta-analytic techniques (e.g. DRINKMANN, 1990; FRICKE & TREINIES, 1985; HEDGES, 1980, 1982a, 1982b, 1983, 1986; HEDGES & OLKIN, 1985; KRAEMER & ANDREWS, 1982; KULIK & KULIK, 1986, 1988; RAUDENBUSH & BRYK, 1985; ROSENTHAL, 1978, 1979; ROSENTHAL & RUBIN, 1982a, 1982b, 1982c, 1986; STRUBE & MILLER, 1986). Reliability of data, another prerequisite of statistical conclusion validity has been the subject of several studies (e.g. COOPER & RIBBLE, 1989; MATT, 1989; STOCK, OKUN et al., 1982). Others call attention to the fact that the statistical techniques employed in meta-analyses may have low power to detect moderators of effect sizes (KEMERY, MOSSHOLDER & DUNLOP, 1989; KEMERY, MOSSHOLDER & ROTH, 1987; SACKETT, HARRIS & ORR, 1986; SPECTOR & LEVINE, 1987). Assuming mistakenly that no moderators exist (Type II error) leads to the inappropriate conclusion that the relation between variables of interest is well understood, whereas the inappropriate search for moderators (Type I error) would be less grave (SACKETT, HARRIS & ORR, 1986).

It seems as though WALBERG's (1984) criticism of largely neglecting threats to this type of validity other than power, reliability and error rate could apply in the case of meta-analyses as well. His list of additional potentially invalidating aspects ranges from the effects of leveling and compositing to problems of units of analysis and incompleteness of variable sets. Some of these issues have, however, been the subject of critical discussions (cf. ORWIN & CORDRAY, 1985).

Leveling and compositing of variables can conceal relations between them and mask their specificity. This links up to the problem of mixing apples and oranges which can, however, also be regarded as being largely a problem of construct validity.

4.2.3 *Construct validity*

Construct validity concerns the extent to which theoretical variables have been successfully operationalized and measured. COOK and CAMPBELL (1979) divided this type into two subcategories. Construct validity of causes refers to the appropriateness with which the theoretical treatments, populations and settings have been translated into experimental procedures and sampling. Construct validity of effects refers to the adequacy with which theoretically intended outcomes can be measured by the techniques employed.

The adequacy of these implementations can only be determined with reference to the theoretical variables involved and the questions being investigated. As before, on the level of the primary studies the problem is largely one of coding the relevant features to allow the investigation of the influence of various operationalizations on effect sizes.

On the level of the meta-analysis the question of construct validity is closely allied to the apples and oranges debate. The sensibility of classifying various implementations of the independent and dependent variables into one category depends on the hypotheses the meta-analysis wishes to test and generalizations intended to be made (cf. section 3.3.1). Using very broad and heterogeneous categories will reduce the specificity of the results by obscuring particular relations and effects. Using narrow definitions will restrict the generalizability. The intentions of the meta-analyst will determine which approach to use (BRYANT & WORTMAN, 1984). For the consumer of the review it is important to know how the individual summarized studies have operationalized the respective concepts, for only then will the reader be able to form an opinion on the scope of the review. This has important implications for the practical relevance of the conclusions reached.

Both construct and external validity are concerned with specifying the contingencies on which a causal relationship depends. According to COOK and CAMPBELL (1979) the main difference between the two is that construct validity is directed at the more theoretical aspects, while external validity is aimed at real target populations of settings, persons and times.

4.2.4 *External validity*

External validity concerns the import of study results for some body of knowledge external to the specific study, i.e. the robustness of its findings (BRINBERG & McGARTH, 1982). Can one generalize to populations, settings, treatments and measurements other than the particular ones used in the studies? Is the substantive

domain adequately represented? Do the findings hold when other methods are employed?

These issues are often obscured on the level of individual studies (WORTMAN & BRYANT, 1985). The meta-analysis can fulfil this crucial task by including studies with diverse samples, diverse operationalizations of theoretical constructs, diverse methodological approaches and settings. It can help identify which implementations are most effective. The heterogeneity indicated, takes the discussion back to the debate of mixing apples and oranges, i.e. aggregating effect sizes over diverse study findings or stratifying according to specific factors to obtain homogeneous subgroups, thus in fact reducing the debate over external validity to a statistical test of the homogeneity of variance (WORTMAN, 1983).

What strategy will serve the meta-analyst best, depends on the questions addressed (BRYANT & WORTMAN, 1984). What the consumer needs to know and understand fully to evaluate the practical relevance of the meta-analysis is the nature and limitations of the research domain reviewed (BULLOCK & SVYANTEK, 1985). The generalization domain is defined by the inclusion criteria, the theoretical content domain in which the hypotheses are tested, details on research design, sample sizes, demographic variables, operationalizations of concepts, in fact by precise information on all coded items, presenting a comprehensive, descriptive picture of all study features and characteristics.

An interesting perspective on the validity of studies also touching the question of practical relevance is the concept 'of prior validities employed by BRINBERG and McGARTH (1982). This concerns the values, criteria or standards used to select elements and relations from substantive domains as important and legitimate objects of investigation. Something similar is implied by what ORWIN and CORDRAY (1985) refer to as quality of macrolevel reporting: the customary reporting practices of a particular domain influencing what variables are conventionally, rarely or never reported or investigated and their customary operationalizations, both possibly limiting the scope of the meta-analysis. The publication policy followed by scientific journals is another related aspect.

Here the boundary between social science and social policy becomes indistinct. Scientific work is influenced by the the unwritten 'laws' of the mainstream of ideology and social politics, fads and dislikes, ideas and 'pet theories' prevalent in the scientific community. Neither the meta-analyst nor the consumer can escape these social influences. They find expression in the studies undertaken and published, consequently also in the research integration. As GOOD (1983b) points out, research does not take place in a vacuum. Researchers study what they perceive as currently important and that for which funding is available.

Whether consumers will find material relevant to their concerns and practical problems will depend on whether these have been examined by the scientific community. Even should they have been, they will have to remember that the variables investigated are not all-encompassing but were selected by scientists working within a specific social context and sphere of influence. As MEEHL (1978, p.807) bitingly formulates: "in soft psychology theories rise and decline, come and go, more as a function of baffled boredom than anything else; and the enterprise shows a disturbing absence of that cumulative character that is so impressive in disciplines like astronomy, molecular biology or genetics." As WALBERG (1984) points out with reference to educational research, the number of possible causes of a given educational effect is indefinitely large. So even if reliability and validity of the meta-analysis seem adequate there is no guarantee that the variables reported represent adequately all those potentially relevant for the substantive domain.

A judgement on the general practical relevance of meta-analyses cannot be made on the basis of personal substantive interests. The fact that meta-analyses do not appear to be available on the topic of primary interest cannot be interpreted as a general deficiency in their practical relevance as such. It is rather an indication that persons conducting meta-analyses have their own specific fields of interest.

If, however, consumers find a meta-analytic study relevant to their interests and are not experts on the approach themselves, they will have to follow some sort of guidelines on how to evaluate the study, else they would not be keeping to the critical stance propagated by the meta-analysts themselves.

5 Guidelines for evaluating meta-analyses

The aim of previously published guidelines is either to help potential meta-analysts to avoid invalidating factors or to present readers with a framework for evaluating the quality of a synthesis, the primary focus of either being the potential meta-analyst as reader.

Apart from specific guidelines, articles frequently allude to minimum requirements needed to be met by meta-analytic research reports (e.g. FRICKE & TREINIES, 1985; HUNTER, SCHMIDT & JACKSON, 1982; JACKSON, 1980; SLAVIN, 1984, 1986; STRUBE & HARTMANN, 1982, 1983). Some feel that possibly the best way to ensure the quality of meta-analyses and increase confidence in the validity of their conclusions is to reanalyze, replicate or criticize them routinely (e.g. BULLOCK & SVYANTEK, 1985; CORDRAY & ORWIN, 1983; FISKE, 1983; GREEN & HALL, 1984; KULIK, 1984; STRUBE & HARTMANN, 1983).

As guidelines represent optimal criteria which meta-analyses should try to meet, and according to COOPER (1982) rarely can attain, they refer both to the aims of meta-analyses and aspects limiting their attainment. The information presented so far is thus also to a great extent the content of guidelines.

COOPER (1982) published one of the first and most frequently mentioned guidelines for conducting integrative research reviews. Based on the conceptualization of integrative reviewing as a research project involving five stages (problem formulation; data collection; data evaluation; analysis and interpretation; public presentation), it discusses ten threats to the validity of meta-analytic conclusions, two for each stage, engendered by the methodological choices a reviewer is required to make. Along with these he presents suggestions on how to protect the validity of the review. Making no claim to completeness and expecting the number of threats identified to increase, he adds one more to the list himself in a later publication (COOPER, 1984).

Rather than focusing on specific stages and potential sources of invalidity associated with each, other guidelines direct the reader's attention more generally toward limitations and pitfalls inherent in the approach in the shape of questions or requirements a meta-analyst should have considered or fulfilled in the process, i.e. what COOPER (1982, 1984) discusses under the heading of protecting validities. These issues can, however, roughly be classified as related to particular stages of the meta-analysis.

Although basically corresponding to Cooper's, the six stages used by JACKSON (1980) in his study analyzing the methodology of reviews seem preferable (selection of

hypotheses; sampling; presenting the characteristics of the primary studies; analyzing primary studies; interpreting results; reporting). The major difference, apart from slight shifts in the scope of the phases, is the splitting of analysis and interpretation into two separate steps. Doing so emphasizes the difference between efforts to reduce data to an interpretable form, enabling one to draw inferences from the studies, distinguishing systematic patterns from noise or chance, and efforts to draw conclusions from these results concerning their implications for policy, theory and future research. This distinction seems sensible considering that one of the aims of meta-analyses is to present summarized research evidence in a practically useful form to all consumers of research (cf. sections 2.1 & 2.2)

LIGHT and PILLEMER (1984) present and discuss a checklist of ten questions for evaluating reviews, applicable to both scientific and policy research. Of these questions one concerns the purpose of the review, two pertain to sampling and selection adequacy, six refer both to coding issues and the subsequent analysis of their influence on beween study variations and finally one to the implications of the review for future research.

BULLOCK and SVYANTEK (1985) developed a list of 14 criteria for evaluating meta-analytic research based on problems encountered during a review and replication attempt of a meta-analysis. Of the criteria, representing quality standards acceptable meta-analyses should meet, two concern the theoretical domain from which and in which hypotheses are formulated and tested, three refer to search and selection strategies, four to coding practices, one to analysis problems, three to interpretational aspects and six in all to reporting practices, five of which overlap with the previous phases.

The latest guidelines are provided by KULIK and KULIK (1989). They formulated six questions, discussed at length in four sections, involving a variety of factors readers should attend to when wanting to distinguish between good and poor meta-analyses. The aspects considered were developed not only to guide their own review of meta-analytic literature, but also for others with similar intentions or persons wishing to conduct their own meta-analyses. Two of the questions concern the location and selection of studies, one refers to coding and three to statistical or quantitative problems.

Evidently these guidelines all cover similar ground and do not present essentially new information. They differ largely with respect to the weight given problems relating to the various stages. All in all coding, sampling and analysis tend to receive most attention. The focus on reporting issues is an indication that the guidelines are addressed primarily to meta-analysts. Another pointer in this direction is that the interpretation or generalization stage is largely neglected.

From the consumer's point of view the whole process of evaluation depends almost entirely on adequate reporting. Unless information is available, there is no way of judging in what manner known limitations or pitfalls have been avoided or handled by the reviewer. Omissions in the meta-analytic report, for example of vital coded data, can be looked upon as not having been taken into account in the integration, for one can hardly expect the reader to go to the trouble of consulting all the primary studies reviewed in order to obtain the required information (as for instance SLAVIN (1984) does in his effort to evaluate the quality of several meta-analyses). Neither should it be necessary to write to the author to be able to critically evaluate the report, for example because detailed information on essential procedural aspects has not been mentioned or the list of the reviewed literature is incomplete or missing. One of the recognized precepts in science is that reports on research should include enough information to allow their critical examination (HUNTER, SCHMIDT & JACKSON, 1982; JACKSON, 1980; MEINEFELD, 1985).

Apart from this, the stand taken here is that from a practitioner's perspective there are specific priorities in evaluating the information contained in a meta-analytic report. It is contended that before going to the trouble of evaluating the meta-analysis as a whole the practitioner should concentrate on the theoretical scope of the study and the domain of generalization, i.e. all matters concerning the construct and external validity of the review. These aspects are related primarily to the stages of problem formulation, sampling and coding.

Should the scope of the study seem to cover the practitioner's interests adequately, the next matter to consider should be the conclusions or implications the reviewer has formulated, i.e. the interpretation stage. If these appear to be practically meaningful and relevant, the consumer will have to attend to the question of internal and statistical conclusion validity. These issues concern the possible presence of bias or confounding variables, the appropriateness of the quantitative or statistical methods employed and the reliability of conclusions. These matters are covered primarily in the coding and analysis stages.

The following recommendations will not present something entirely novel, but accentuate different aspects and try to suggest ways of circumventing expert knowledge where this should be necessary, as for example when trying to evaluate the adequacy of specific statistical techniques or research designs. They do not lay claim to comprehensiveness. They are the result of personal experiences made in working with meta-analyses and should be looked upon as procedural tips rather than rigid principles to be followed unwaveringly in all cases. The individual reasons for consulting a meta-analysis will largely determine which points receive primary attention.

5.1 *Recommendations on how to evaluate a meta-analysis*

As indicated above, the evaluation process is subdivided roughly into two phases. The first concerns the practical scope of the meta-analysis while the second refers to what is usually conceived of as being the actual examination of the quality of a study: whether the relations identified are so for the assumed theoretical reasons or whether some other factors might be responsible for the effect. Only if consumers decide that the study is practically useful to them does a detailed analysis of the review's quality seem necessary, using the aims and limitations meta-analysts themselves have formulated for the approach as general guiding principles, i.e. applying the information summarized in the previous sections.

Assuming that readers consult a meta-analysis with specific interests in mind, it follows that determining what exactly the review covers is foremost on their minds. It seems sensible, therefore, to begin the critical evaluation with aspects that help clarify the extent to which these interests and the scope of the study coincide. By first concentrating on the construct and external validity of the meta-analysis (cf. sections 4.2.3 and 4.2.4), one will at the same time be able to form an opinion on its practical relevance for oneself, thus disposing of two tasks simultaneously. Furthermore, should either the quality of these aspects or the degree of overlap be found wanting, one can decide to break off the evaluation process, saving time and effort. If, however, readers decide their interests are adequately represented, they will have to analyze the quality of the meta-analysis with respect to internal and statistical conclusion validity (cf. sections 4.2.1 and 4.2.2) to ensure that the results and conclusions are trustworthy, valid and reliable.

The first phase of the evaluation can be conceptualized as a screening process, identifying whether the particular meta-analysis seems relevant to one's own purposes, thus worth the trouble of the second phase of evaluating the quality in detail. More specifically this means that one is first concerned with the variables investigated and whether subjects and settings in the primary studies correspond with one's own target populations. The second phase involves the comparison of the meta-analysis with the rigorous scientific standards the experts have set themselves. This implies that the basic requirement is intersubjective testability: a replication of the meta-analysis should be possible using the data presented in the report. Furthermore, it means that reviewers should show definite efforts to ensure the reliability and validity of their work by taking a critical and problem orientated attitude toward what they are doing. The more this self-critical stance is evident, the less details and problems are glossed over in all the stages of the research integration, the more confidence one can have in the quality of the meta-analysis.

Both phases are interrelated and concern the validity and reliability of the meta-analysis as a whole. While going through the screening phase one will inevitably be con-

fronted with the question of the quality of these aspects. Already evaluating them during this phase enables one to refer back to these judgements when in the second phase one tries to form an overall estimate of the quality of the meta-analysis.

5.2 *Construct and external validity*

Where in the meta-analytic report will one find the information relevant to the theoretical scope of the study and its generalization domain? What should be looked out for to evaluate its quality in this respect? Theoretically the stages involving problem formulation, hypothesis selection, coding and interpretation are concerned. Practically, however, the information is scattered throughout the report in diverse passages, figures and tables. A certain amount of detective work will usually be necessary to glean the relevant data from it. There is usually no way to avoid reading the complete report, unless its aim is obviously not what was hoped for with only the title or abstract to go by.

It simplifies matters if, while reading, one keeps in mind the issues that help delineate the scope of the study and determine the quality. Apart from needing the definitions of the theoretical variables and their operationalizations, one has to determine the persons, settings and times covered by the reviewed studies. For the sake of clarity the aspects will be discussed under the heading of the various reviewing stages, even though one might not find an exact counterpart of this structure in the actual meta-analytic report.

5.2.1 *Problem formulation and hypothesis selection*

The aims formulated by meta-analysts and the hypotheses they intend testing give the first indication of the theoretical scope of the study. How can a non-expert judge the quality of this information? A workable approach is to analyze how critically and problem orientated the reviewer has gone to work.

Attention to controversial issues in the field, mention of rival or contradictory theories or findings, discussion of possible confounding, intervening or moderating variables are indications of such an attitude. Relating the specific aims and hypotheses of the meta-analysis to these issues, explaining why these and not others were chosen, thereby placing the study in the general context of prior theoretical discussions, reviews or influential primary research, should enhance confidence in the theoretical framework outlined.

If no attempt is made to explicitly define the theoretical construct variables relevant to the aims and hypotheses, readers should try to do so themselves, using all the informa-

tion presented in the report, e.g. in the formulation of the selection criteria, by inspecting the coded items, or details mentioned in the discussion section of the report. Usually, one gains some impression of the theoretical framework the meta-analyst had in mind. More often than not, however, instead of theoretical definitions one discovers the operationalizations of the relevant variables, thus an indirect indication of how the reviewer might have conceptualized them. So for example, instead of a description of what theoretical models of intelligence are examined in the individual studies and what this implies, one might find a list of the tests used to measure this abstract construct, giving some impression of its conceptualization.

5.2.2 *Sampling and selection of studies*

Both the search strategy and the selection criteria give one a more concrete idea of the theoretical scope. The quality of these procedures depends to a large extent on the quality of the theoretical framework developed.

Having heeded the diverse theoretical positions held in the domain, the terminology used to locate relevant studies will have a greater likelihood of leading to a representative sample. Confidence in the quality of the search is increased if various sources are consulted to find studies, if details of the procedure are presented, if possible sources of bias are discussed and taken into account (cf. sections 3.2.1 and 3.2.2). One should gain the impression that everything was done to ensure an exhaustive sample, leaving no stone unturned to gain access to as much published and unpublished research as possible, e.g. searching various databases such as ERIC, Psychological Abstracts or Dissertation Abstracts, consulting diverse bibliographies or contacting experts in the field for information on relevant studies.

A more definite delimitation of the generalization domain is obtained by studying the inclusion and exclusion criteria. Both are equally important for clarifying the actual scope and should thus be analyzed in detail. Once again, one can assume that the better the theoretical understanding of the domain, the more likely the selection criteria formulated will ensure the appropriateness of studies.

Neither breadth nor narrowness of the selection procedure per se increase or reduce the validity of the meta-analysis. The critical question is how the criteria employed correspond with the purpose of the study. A restrictive database for meta-analyses aiming at broad generalizations or being largely exploratory in character would not be appropriate. Neither would a broad database covering diverse theoretical variables and their operationalizations be adequate for meta-analyses wishing to investigate a theoretically very specifically delineated aspect.

What is essential is that the criteria used be clearly specified so that a comparison with the meta-analytic aims is possible. The reader should be able to identify the range of theoretical features and actual characteristics of the studies included as well as excluded from the integration, as this explicitly defines the content and generalization domain.

A self-critical attitude on the part of the reviewer, attention to details, limitations and aims of the strategies can serve to enhance the confidence in these procedures. Should one doubt the thoroughness of the search or the adequacy of the selection, the results should be handled with care.

More often than not a clear picture of the actual theoretical and practical scope of the study emerges only when examining exactly which aspects were coded and eventually analyzed in the review.

5.2.3 *Presenting the characteristics of the primary studies*

Not all the features coded pertain directly to issues of construct and external validity. The items of primary interest are those referring to the variables studied and the characteristics of persons and settings.

One needs to gain a comprehensive idea of how variables were translated into practice in the individual studies. As before details are essential, e.g.: How were the theoretical variables operationalized or measured? For how long were interventions applied? At what stages were the outcome variables measured? Which moderating or intervening factors were studied? How were the various experimental and control groups defined? Which of these distinctive study features did the reviewer classify as belonging to the same category for the purpose of statistical analysis? Does this process blur or gloss over important theoretical distinctions? The data should cover all aspects serving to define the construct variables. By comparing these implementations with the actual theoretical definitions formulated one can make a rough estimate of their correspondence, i.e. the degree to which the given data would possibly allow the hypotheses to be tested adequately.

Equally important for the practitioner are the characteristics of persons, settings and times covered in the studies. Do they correspond with the target populations one has in mind? The data should reveal all theoretically relevant aspects concerning the persons examined, e.g. age, sex, diverse personality, demographic or educational variables. The same applies to the settings, e.g. field or laboratory investigations, with existing or specifically assembled groups, in restricted or broad geographic regions, published over a long time period or concentrating mainly on fairly recent investigations.

These aspects should all help to clarify how close to the practical reality of the reader the reviewed studies are. Once more confidence in the ability of the coded items to transmit an adequate picture of the distinctive features of the studies should be enhanced if the reviewer developed the framework for this on the basis of the theoretical background involved, remaining flexible enough to attend to important distinctions discovered while analyzing the individual studies, covers moderating variables known or suspected of causing possible interaction effects and discusses qualitatively those issues not coded.

After this detailed analysis readers should be in a position to decide whether the review corresponds with their personal interests. Should this not be the case, one can put the article aside. Otherwise, the next step in the screening phase of the evaluation process is a critical examination of the conclusions presented in the interpretation stage of the meta-analysis.

5.2.4 *Interpreting results*

Ideally, one would expect the reviewer to specify the generalization domain to which the conclusions can be applied. This is, however, rarely the case. One will usually have to rely on the description of study features, both theoretical and practical, consider the adequacy of the sample of individual studies and decide for oneself whether the results can be transferred to the particular situation one has in mind.

What the consumer has to evaluate at this stage is the interpretation of effect sizes and results as well as how these are presented. Is one simply confronted with quantitative results, left to give these meaning oneself or does the reviewer attempt to interpret them, referring to their practical and theoretical implications? Are the conclusions formulated with regard to their application domain, indicating possible mediating or intervening factors restricting or delineating their applicability? Is the meaning of the effect sizes explained and the range of studies and variable operationalizations to which they pertain or must one figure this out for oneself?

The conclusions drawn from the results should not be formulated in a causal way. A meta-analysis does not allow this type of interpretation. What it performs is a quantitative description of the general tendency of findings derived from a comprehensive sample of research evidence on a specific topic.

The more attention to detail and the more references to the particular theoretical and practical relevance, the more confidence one may have in the quality of the interpretation and the sensitivity with which the meta-analyst has gone to work. If, however, results are presented primarily in quantitative form, with little attention to interpretation

and implications, one might feel that the reviewer aims to impress with numbers and significance levels that actually might only have little practical or theoretical relevance.

Meta-analysis aims at practical simplicity and comprehensiveness (cf. section 2.1). The conclusions should thus neither be obscurely quantitative, understood only by the initiated, nor simplistic in the sense of ignoring the complexities involved in the reviewed domain.

Should the conclusions not be presented by the reviewer in a form that can be made practical use of, consumers can try to interpret the results themselves, using the information gathered in the process of evaluating the theoretical content and generalization domains. In either case, if the resulting conclusions seem to answer the questions one had in mind when consulting the meta-analysis, one will have to face the more ticklish problem of evaluating the internal and statistical conclusion validity of the study.

5.3 Internal and statistical conclusion validity

Were the study designs and statistical techniques employed in the primary studies and the meta-analysis itself likely to yield results that allow valid conclusions or could some other factors possibly explain the obtained results? How does one evaluate the design quality or the adequacy of statistical methods without expert knowledge? On the level of the primary studies one will have to rely largely on the information presented by the meta-analyst, especially in the coding stage. On the meta-analytic level the consumer can form an independent opinion by concentrating mainly on the information given in the analysis stage while additionally referring back to quality judgements made in the previous phase of the evaluation process, trying to decide whether a replication of the review could be possible with the information given.

5.3.1 Coding

As there are no absolute quality standards, the reviewer's sensitivity to and awareness of the problems concerned will have to serve as indication. If this is not available in the text of the meta-analytic report, the alternative of analyzing the coded study features exists.

The more aspects relating to the validity of the primary studies are coded, the more likelihood that the meta-analyst will be in a position to examine their possible influence on between study variation of effect sizes. A rough estimate of the quality of this information can be made by trying to judge whether the data would allow one to reconstruct study differences. The least one would expect the coded data to cover is: sample size,

unit of analysis, assignment rules and study design. Furthermore, some indication of the adequacy of variable implementations, statistical techniques and reliability of outcome measures should be given.

Apart from this, as coding represents measurement in meta-analyses, details on the reliability of the process are essential to be able to evaluate the quality of the meta-analysis itself. Were several coders, preferably blind to the aims of the meta-analysis and the results of the primary studies, employed? Are coding reliability or interrater agreement coefficients reported for the individual items or rather just averages? Are problems concerning the reporting accuracy of primary studies, difficulties because of missing data or errors and how these were handled, mentioned? No one can extract valid and reliable results from faulty data. The reliability of coding is therefore of great importance. Should one have serious doubts in this respect, it might be sensible to discontinue the evaluation.

Computing the effect sizes is another crucial part of the coding process. As experts point out, there are various formula available, leading to different values of effect sizes. Misapplication, especially if studies used diverse designs, can lead to artifactual between study variations. One should thus examine how the formula were employed. If the review covers studies with different experimental designs and the reviewer has not analyzed their influence on effect size variations one should inspect the relevant data oneself. If this is not possible, the least one should find is an exact report on the computation procedure so that other experts would potentially be able to critically evaluate the techniques used.

With the appraisal of effect size quantification one has taken the first step toward evaluating the statistical methodology employed in the meta-analysis. Because experts do not agree as to which approaches are most appropriate, one should try to find critical evaluations or replications of the study to discover how specialists rate it. Additionally, these often confront one with controversial points that had eluded one's attention.

5.3.2 Statistical analysis

A large portion of the statistical analysis should be devoted to efforts on the part of the meta-analyst to establish the validity of the review. This primarily concerns the search for moderating variables and analyzing the heterogeneity of effect sizes.

For example, has sampling bias been avoided? This can be examined by studying publication source as possible factor influencing effect sizes. If such an effect is present and the meta-analysis relies primarily on one source of studies, generalizing the results becomes problematic. Similarly, if publication date or quality aspects of studies seem to influence between study variations in effect sizes, one will have to take extreme care

when generalizing results, especially if the reviewer presents these as averages computed across heterogeneous studies.

Generally, before computing aggregated effect sizes, the homogeneity should have been investigated. If not, did the reviewer have specific reasons for this, e.g. broad questions? Or are these global averages presented along with others for more homogeneous subgoups? If the relevant information is available, one should inspect the range of effect sizes oneself. If outliers or distribution irregularities are present, one will have to try and identify which of the coded factors could possibly be responsible for these variations. At any rate, one will have to be careful about the domain one can generalize to.

Apart from these moderating or mediating influences, the unit of analysis employed in the meta-analysis is a critical issue in evaluating the adequacy of the results. Whereas in primary studies using different units of analysis can lead to artifactual variations in effect sizes, the units of analysis employed in the meta-analysis can cause the problem of non-independent data, making the statistical analysis itself problematic (cf. section 3.3.2). If study findings and not the studies themselves were used as unit of analysis, how was the problem of non-independence handled? Was it ignored or taken into account by weighting or some other procedure? An additional source of non-independence can occur, if the reviewed sample contains several studies by the same author or affiliated groups of researchers. If the reviewer does not examine this aspect, one should try to clarify the point oneself, e.g. by studying the names of the authors of the articles being integrated. A question one will have to answer is whether this represents a form of selection bias or whether the topic of interest is really predominantly investigated by this group of scientists.

As more than one technique is available for analyzing the same question, confidence in the results is increased, if these lead to similar conclusions. Should parallel analyses not have been conducted, the reviewer should at least have examined whether the assumptions for the valid application of the techniques are met by the data, especially if parametric methods such as ANOVA, regression analysis or t-tests were employed. If not, does the reviewer at least point to possible limitations of these techniques?

If several significance test were performed, does the reviewer enter into the problem of capitalization on chance or refer to the effect a large review sample size might have on the significance of results? On the other hand, if relatively small samples of studies were used, was the file drawer problem examined or the issue of representativeness discussed?

In general, if meta-analysts take their claim to rigorous scientific standards seriously, they should show this in a critical attitude toward their own work. Should this not be the case nor known limitations or problems of the approach be the subject of critical discussion, one should be sceptical of the results. No study is without flaws or problems, but

these should at least be pointed out. By formulating rigorous standards and exposing the limitations of their approach, meta-analysts have put themselves in a position inviting critical evaluation. The consumer should thus follow their pointers in these directions and tick off which of the identified problems have been dealt with and which ignored. The more of these potential limitations and demanded standards are neglected, the more sceptical one should be of the validity and reliability of the results.

Those aspects that need expert knowledge to be evaluated in detail can be circumvented by sticking to these more superficial and attitudinal issues. It is advisable to search for expert appraisals or replications of the meta-analysis. On the one hand these may uncover crucial aspects one could not evaluate oneself, but on the other hand also identify points that on secondary analysis, contrary to one's own or other evaluations, do not appear to critically jeopardize the validity of the study.

5.4 Of what use is the evaluation?

After this analysis one is faced with the problem of deciding whether the quality of the conclusions reached is sufficient to allow their transfer into practice. This is a difficult problem. How good is good enough, especially as evidently no study is without some flaws or problems? More often than not the decision will be an entirely subjective one, in which the consumer weighs the gravity of the quality deficits against the problems requiring a practical solution and the possible negative effects or costs one would incur by applying measures that might not live up to one's expectations.

Whatever final judgement the evaluation eventually leads to, having made a detailed analysis will help in formulating reasons to justify it. Even if the validity and reliability of the meta-analysis are not satisfactory, the review is never entirely worthless for it can always serve as bibliographic source, should the reviewer have included a list of the articles integrated.

The evaluation procedure presented above was developed in the course of working with meta-analyses on specific educational topics for the purpose of finding well founded empirical research evidence to help substantiate relations hypothesized in a complex theoretical network of educational variables concerning motivational and disciplinary problems in secondary school. The following section attempts a summary of the impressions made by the meta-analyses read and evaluated according to a slightly adapted version of the procedure presented above.

6 Evaluation of a sample of meta-analyses

6.1 Method

6.1.1 Sample description

The sample of meta-analyses summarized in this section is more or less a chance collection resulting from a dual search strategy: finding either studies relevant to the educational variables of the theoretical network mentioned above or articles potentially helpful in coming to grips with methodological or evaluation problems. In either case the bibliographies of the books by GLASS, McGAW and SMITH (1981) and FRICKE and TREINIES (1985) served as starting-point. Besides this, the databases PSYINDEX and PSYCINFO were searched from 1978 to 1988 for meta-analytic studies in the educational field using the descriptors: meta-analysis, integrative review, data synthesis, quantitative assessment and statistical review. The bibliography of every article read was inspected for references to other relevant papers. Furthermore, the current content of various educational research journals was monitored in the hope of finding additional studies, an effort that is still continuing.

Initially, meta-analyses were only evaluated in detail, narratively, if they covered aspects relevant to the theoretical network. In this event critiques or replication attempts were looked out for. However, to place the present assessment on a broader footing all the meta-analyses read in the course of our work were included in the evaluation.

In this way a total of 55 articles was obtained, 48 of which were considered 'independent'. The 'dependent' studies used the same literature base to analyze either different outcome measures (GLASS & SMITH, 1979; SMITH & GLASS, 1980; STOCK et al., 1983; WITTER et al., 1984) or different methodological aspects of the meta-analytic process (BRYANT & WORTMAN, 1984; STOCK et al., 1982); or were publications of apparently the same meta-analysis in different journals or books (FRICKE, 1985; FRICKE & TREINIES, 1985; KLAUER, 1981, 1984; RAUDENBUSH, 1983, 1984). Figure 1 shows the distribution of the evaluated studies per year of publication. The articles referred to by the various numbers are listed in Appendix 1 which also contains references to critiques or replication attempts. Those with the same number but appended letters are dependent studies.

Figure 1. Distribution of evaluated studies per year of publication (n=55)

12					*48*					
11					45					
10			43	47	44					
9			*41*	*46*	*42*	39				
8			*34*	*36*	*38*	31				
7			*23*	*35*	*37*	28b				
6			18	*32*	30	25c				
5			*16*	25a	25b	*22b*				
4		*33*	*13a*	*19*	22a	20	28a			
3	26	*29*	*12*	10	8	15	27			
2	*17*	21	*9*	*3*	5	14	6b		24	
1	*7a*	*7b*	*2*	1	*4*	13b	6a		11	40
year	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988

Numbers refer to the article code (cf. Appendix 1)
Italics refer to the 25 studies also coded by Fricke & Treinies (1985)

6.1.2 Coding procedure

To avoid basing the summary on purely subjective impressions a rough coding scheme was devised on the basis of the methodological issues discussed in the previous sections, covering aspects concerning the theoretical framework, sampling, coded characteristics, data-analysis and interpretation (a total of 80 categories, cf. Table 1). The meta-analyses were coded independently by a colleague, Andrea Mertens (Frankfurt), and the author, hoping to ensure the quality of the database by checking the intercoder agreement (critiques or replication attempts were not coded or analyzed).

Following this procedure a maximum of 4400 aspects could have been coded. In all only 2093 (48%) recordings were made, reaching an overall intercoder agreement of about 88%, even though the coding system was not very sophisticated. These differences were subsequently discussed to achieve a mutually acceptable classification. Discrepancies occurred mainly because information in the article had been overlooked or something had been coded which on discussion turned out not to represent the category adequately after all. Lack of clear distinctions and a certain amount of overlap between some of the categories also contributed to discrepant coding. In a number of cases the information in the article was either too ambiguous to allow certainty in coding or only partially fitted the category. Nonetheless, taking a lenient stance, this information was coded, but placed in parenthesis (cf. Appendix 2 for the coding tables).

6.1.3 *Data evaluation*

A lenient approach was also taken in computing the frequency of recordings per category. Thus, if the information needed to code the category could be extracted from any one of the 'dependent' articles, it was counted as present for this meta-analysis, even though it might not have been mentioned in every one of the publications. Essentially, this means that either two or three articles were looked upon as a single study, thereby increasing the likelihood of obtaining the information. The maximum frequency possible per coded item was thus 48, i.e. the number of 'independent' meta-analyses.

It is possible to characterize how meta-analysts typically conduct the various stages of the meta-analytic process on the basis of these frequencies. They show how the identified aims and problems are being handled and thus provide a rough impression of the general quality of meta-analyses.

Additionally, the empirical database will be compared with two assessments of narrative reviews made previously. This indirectly allows one to estimate whether or not meta-analyses are improving the quality of integrating empirical research in general. JACKSON (1980) analyzed a random sample of 36 reviews using a 66 item coding instrument. His purpose was to examine the methods used and their frequency, to evaluate them critically and suggest more powerful ways of integrating research findings. WAXMAN and WALBERG (1982) assessed a systematically collected sample of 19 narrative reviews on the process-product paradigm published between 1970 and 1979, one of their aims being to help improve the quality of future reviews.

The database will also be compared to the results obtained by FRICKE and TREINIES (1985) who examined the statistical validity of a sample of 67 meta-analyses. Even though their study only covers part of the issues coded and analyzed in the present study, the comparison will give some indication as to whether the identified trends are specific to the present sample or can possibly be generalized. The coding categories are not absolutely equivalent, yet correspond enough to justify this procedure. Furthermore, as FRICKE and TREINIES (1985) reported their coded items per study and 25 of their meta-analyses are also included in the present sample (cf. Figure 1, italics) an indirect check on the coding reliability could be made for certain categories. Most of the recordings were similar. The differences will be referred to in the discussion of the relevant results.

6.2 *Results*

One of the things noticed during the coding process was that the distinctive subjective impressions gained by the raters while reading the individual articles was lost. This

might to a certain degree have been caused by not using graded rating scales in the coding scheme to document the variations in depth and quality. Additionally, no categories had been included to cover such matters as the readability of the text, the clarity of its presentation, the frustration of having to page back and forth to find the relevant information or doubts raised concerning the quality of the work because of errors noticed on closer reading of the text, tables or bibliographies, often indicating a lack of care.

These subjective impressions will be used to help colour the discussion of the results summarized in Table 1. Furthermore, if possible the findings will be contrasted with those of the three studies mentioned above (cf. p.51). To ease the interpretation and understanding of the results, it should be noted that the percentages presented per coded item cannot be added across individual items to add up to a 100%. Even though some of the categories appear to be mutually exclusive, the nature of both the meta-analyses and the coding procedure allow a study to be coded in each of them. The reason for this is that meta-analyses concern complex theoretical frameworks, many variables, methods and analyses. For some of these the information necessary for coding is available, whereas for others it is not. Additionally, while one of the articles making up an 'independent' meta-analysis may contain the information, another might not. Thus, the percentage per coded item has to be looked at individually, based on a maximum of 48 possible recordings, as noted above.

Table 1. Frequency and percentage of records per category (n=48)

		n	(n)*	%
I	Theoretical framework			
1	-problem orientated	38		79
	-indirectly deducible	14		29
2	-prior reviews mentioned	43		90
	-prior research mentioned	34		71
	-shortcomings indicated	27	2	56
3	Construct definitions			
	-specifically formulated	38		79
	-indirectly deducible	28		58
	-diversity ignored	3		6
	-diversity analyzed	35	2	73
4	Aims			
	-exploratory	46	2	96
	-specific hypotheses	7	1	15
II	Sampling			
1	Search strategy			
	-retrieval system	38	2	79
	-bibliographies	34	1	71
	-experts quizzed	3		6

Table 1 continued		n	(n)*	%
2	Representativeness			
	-published/unpublished	27		56
	-file drawer/fail safe	9	4	19
	-source variation	23	1	48
3	Selection criteria			
	-specified	34	1	71
	-excluded studies	15	3	31
	-articles listed	31	4	65
	-available at request	13	1	27
4	Number of studies	median=43	range: 9-153	
III	Coded characteristics			
1	sample size	20	9	42
2	person characteristics			
	-sex	21	2	44
	-age/class/grade	42	2	88
	-educational variables	25	1	52
	-demographic variables	20	1	42
3	settings			
	-laboratory	12	2	25
	-field	31	6	65
	-region	15	1	31
	-publication date	37		77
4	study features			
	-global quality index	14	2	29
	-sampling/assignment	22	1	46
	-operationalizations	45	1	94
	-reliability of outcome	30	8	63
	-interaction/control var.	19	1	40
	-unit of analysis	6		13
	-adequacy of stat. analyses	10		21
IV	Data analysis			
1	Coding quality			
	-missing data problem noted	32		67
	-information gained elsewhere	8	2	17
	-studies excluded	18	1	38
	-global intercoder agreement	12	1	25
	-indices per item	2	1	4
	-strategy to reach agreement	9		19
2	Effect size conceptualization			
	-formula/index indicated	47	2	98
	-magnitudes listed per study	17	4	35
	-stem-leaf/graph summary	23	6	48
	-significance indicated	12	5	25
	-outcome variable indicated	29	7	60

Table 1 continued		n	(n)*	%
3	Unit of analysis			
	-study	19	5	40
	-study findings	31	3	65
4	Non-independence			
	-present,taken into account	32	6	67
	-ignored, authors	28	1	58
	-ignored, findings	25	6	52
5	Average effect sizes for			
	-total sample	27	2	56
	-subsample	47		98
	-range/std.dev./std.error	43	4	90
6	Statistical analysis			
	-conventional	36	4	75
	-'modern'	12		25
	-assumptions tested	4	1	8
	-limitations noted	13	1	27
	-parallel analysis	14	1	29
	-capitalization on chance	4		8
7	Effect size variability			
	-heterogeneity tested	11		23
	-quality aspects	24	2	50
	-treatment variation	39	3	81
	-outcome variation	43	8	90
	-subject characteristics	38	2	79
	-design/stat.analysis	20		42
	-contextual/scope variables	39		81
V	Interpretation			
1	Effect size			
	-std.dev/percentile	34	1	71
	-binomial effect size display	1		2
	-Cohen's classification	14		29
	-behavioral indices	5	1	10
	-expert judgement	1	1	2
2	Theoretical implications			
	-old theory/impressions	40	2	83
	-new theory/hypotheses	5	4	10
3	Practical implications			
	-for policy or practice	18	2	38
	-limitations noted	34	2	71
4	Future implications			
	-for primary research	32	2	67
	-for reviews	12	2	25

* the number of uncertain or partially fitting recordings contained in the total frequency of the item

6.2.1 *Theoretical framework*

In 79% of the meta-analyses authors made some sort of an attempt at a problem orientated introduction to their work. Although this appears fairly high, the percentage disguises the actual variations in quality or depth of the presented information. The majority place their study into a rather general historical perspective by indicating research trends in the field and mentioning that findings have been inconclusive without, however, going into a detailed description of the prevalent theories or methodological problems of the domain. These matters are usually covered in a relatively short section of the report, the bulk of the article being devoted to meta-analytic methodology, results and analyses, often leaving the reader with only a superficial idea of the substantive issues involved. This impression is underlined by the fact that in 29% of the cases aspects of the theoretical background had to be deduced indirectly by studying the selection criteria, coding categories or the discussion of the results and conclusions.

A similar picture emerges when analyzing the number of previous reviews or research studies mentioned in the theoretical introductions. Even though these are presented in 90% or respectively 71% of the meta-analyses, they are rarely described in detail. As above, the qualitative differences are not reflected by the percentages. The same applies to the shortcomings noted: these were pointed out, predominantly in a very general manner, for either reviews or research in 56% of the sample. JACKSON (1980) reports that 75% of his sample mention previous reviews, only two of these 27 studies providing a critique. Thus, in comparison, the situation seems to have improved somewhat. However, one should remember that only a minority of the meta-analyses give in-depth descriptions or critiques of the reviews.

An important aspect of the theoretical framework is the definition of construct variables. Seventeen of the 48 meta-analyses (35%) contained both well and inadequately defined variables. Whereas 79% of the meta-analyses present some or all of the constructs with explicit definitions or adequately clarify their meaning through theoretical background information, in 58% of the meta-analyses it proved necessary to deduce either some or all of the variable definitions indirectly, using the coding categories or vague statements. Once again, there is a wide range in the quality of these definitions. Some present details, reporting the actual instruments or implementations used, others indicate the variety of operationalizations through some sort of enumeration, the majority, however, lack this detail, leaving the reader with just a hazy impression of diversity.

Only 6% of the meta-analyses explicitly report ignoring the diversity in their statistical analyses. A total of 73% take it into account in varying degrees of differentiation, most using a variety of fairly broad treatment or outcome variable subcategories, a notable ex-

ception being the meta-analysis by HANSFORD and HATTIE (1982) who examined a wide range of operationalizations.

The trend to keeping matters fairly general also emerged when looking at the aims formulated for conducting the meta-analyses: whereas 96% were exploratory in character, only 15% used the theoretical background to develop more specific hypotheses.

The overall impression created in evaluating the theoretical framework of meta-analyses was that although most authors seem to make an effort to give readers an adequate informational background, they are often left with more questions than the content presented can answer. The quality of meta-analyses would improve considerably, if more detailed substantive information were delivered and this were presented in a slightly more organized fashion, perhaps emulating the structural format commonly used for reporting empirical research.

6.2.2 Sampling

The sampling process of the reviews analyzed by JACKSON (1980) was ill-defined: only one reported the information retrieval system used and three mentioned trying to locate studies via bibliographies of previous reviews. Similarly, WAXMAN and WALBERG (1982) established that none discussed their search procedure. This situation seems to have changed dramatically: 79% of the meta-analyses report the index or retrieval system used in their search and 71% utilized the bibliographies of either prior reviews or research studies. The quality of this information varies considerably, a fact not made evident by the percentages. Some note the databases and years searched, listing the descriptors used. Others only present the name of the indexes without further detail. The suggestion to quiz experts in the field to ensure locating a representative sample was followed by three meta-analysts (6%) only.

Although matters have improved, this progress appears in a different light if one recalls that meta-analysts wish to adhere to rigorous scientific standards, allowing others to replicate their work. The general impression was that only a minority of the meta-analyses would fulfil the criterion of replicability as measure of quality in this respect. It seems doubtful whether other researchers could locate a similar set of studies with just the information in the article to go by. Most reports do not appear precise enough.

Similarly, disregarding their demanded standards, only about half of the sample considers questions of representativeness. Although 56% report including both published and unpublished papers, 48% analyzing this as possible source of effect size variation, only 19% make concrete statements concerning the actual representativeness of their sample. This was done narratively in 9% of the cases, while 10% calculated a fail-safe-

test or file drawer index. The latter corresponds to a certain degree with the 15% FRICKE and TREINIES (1985) report as having computed a fail-safe-test in their sample of 67 meta-analyses (the coding of the 25 identical studies corresponds). As mentioned in section 3.2.1 doubts have been raised about the statistical adequacy of the technique, so that the small percentage might even be interpreted as a positive sign. Nonetheless, it had been expected that meta-analysts would attend to this issue more emphatically, even if only in a qualitative manner.

The selection criteria for studies were specified in 71% of the meta-analyses. These, by implication, often indicated what was not included in the sample, e.g. seriously flawed studies or those for which effect sizes could not be calculated. However, only 31% attempt to describe the excluded studies explicitly, most frequently by reporting an exemplary one. Detailed analysis of rejected studies is rare, the notable exception in this sample being the work of BRYANT and WORTMAN (1984). Nonetheless, meta-analysis seems to have improved matters. In their evaluation of 19 reviews WAXMAN and WALBERG (1982) report that only one explicitly lists the inclusion and exclusion criteria, many of the others are vague, discrepant or arbitrary and three mention no selection criteria at all.

In 65% of the meta-analyses a list of the articles was published, in 27% it was available at request from the authors. This means that in 35% of the cases the reader cannot ascertain what kind of studies the meta-analysis is based on, quite apart from the fact that some of the lists were incomplete or appeared faulty.

In sum, the information presented by the majority of the meta-analyses makes it very difficult for readers to judge the representativeness of the sample by themselves, especially as the number of studies excluded often exceeds the number finally selected. Additional uncertainty is created by the manner in which some of the sample sizes are described. While some authors count the number of articles, others additionally state the number of 'studies', i.e. subdivide an article into separate studies because various independent samples were analyzed. Others present a sample size for their meta-analysis which encompasses all the articles 'meta-analyzed' whereas in reality the analyses are conducted on much smaller subsamples of studies. So the reader is often left wondering, what the actual sample size is. One wonders whether all this play with numbers is necessary. It would simplify matters, if authors were more specific in this respect. Another irritating impression gained was that in some cases the numbers noted did not tally with what was presented in tables or references.

This numerical uncertainty was noted in about 17% of the meta-analyses. Despite this confusion, the median and average sample size was calculated for all articles analyzed. Ranging from 9 to 153 studies, the median was 43, the average 53, indicating a distribution slightly skewed toward the smaller sample sizes.

To summarize the overall impression concerning the quality of the sampling process: although matters seem to have improved, further improvement is possible and necessary to fulfil the standards meta-analysts have set themselves.

6.2.3 Coded study characteristics

As most meta-analyses do not contain narrative descriptions of the primary studies, the coded information is one of the sources from which readers can determine their scope (the populations, settings and span of time covered) and specific features or quality aspects.

On the one hand coding the sample sizes of the primary studies clarifies the number of persons the meta-analytic conclusions are based on, on the other gives some indication of how well founded the individual effect sizes are. While 42% report the approximate number of persons studied, only 23% of the meta-analyses present this information for each primary study. This corresponds fairly well to the 22% found by FRICKE and TREINIES (1985). There were no differences in the coding of the 25 identical studies.

The characteristic of the sample population recorded most frequently is the age, grade or class level of the subjects (88%). Educational or personality variables such as previous achievement or intelligence were coded in 52% of the cases, whereas the sex of the subjects was coded by 44% of the sample. Demographic variables such as socio-economic status or ethnicity were reported by 42% of the sample.

The study settings are usually not presented in as detailed a manner as the person characteristics. Whether the primary studies were conducted in artificial, laboratory settings or whether natural, field studies were carried out, was often not coded or explicitly described, but rather mentioned in the specification of the selection criteria or had to be deduced from the nature of the data presented. The frequencies obtained in this way show that in the present sample field studies predominate (65%), 25% having been carried out under artificial conditions. However, as about half of the sample neither report the information nor could be coded with certainty, practitioners would be hard put to decide whether the results can be transferred to their field of action.

A similar situation exists with regard to the regions in which the studies were conducted: 31% report either the states, countries or specific areas in which the study was carried out or the origins of their sample. In contrast, the span of time covered by the research examined in the meta-analysis was mentioned by 77% of the sample, either by indicating the publication dates or the years included in the search for studies. Although not specifically coded in the present sample, the time factor is seldom analyzed as possible source of effect size variation.

The study feature coded most often was the operationalization of variables (94%). Not expressed by this percentage is the extreme variation in detail. As noted above, during the discussion of the construct definitions (cf. p.55), some present very exact and differentiated information while most give rather global descriptions, leaving the readers to fill in the details by intuition or imagination, perhaps feeling that common sense will suffice to determine the meaning or considering it self-evident.

WAXMAN and WALBERG (1982) analyzed the extent to which the 19 reviews mention any of the 33 threats to validity listed by COOK and CAMPBELL (1979). They report that 95% of the codings were cases of ignoring these specific threats to validity. After reducing the number of categories to 12, the rate of coverage was raised to 15%, varying from the mere mentioning of threats, to illuminating them with a few studies or rarely by a comprehensive coding of all studies. On the other hand JACKSON (1980) reports that 26 of the 36 reviewers (72%) described what were considered major methodological difficulties or shortcomings of the studies reviewed. This kind of information is rare in meta-analyses. However, by coding various aspects indicative of the quality of the primary studies, similar ground is covered.

In 29% of the meta-analyses a global index of quality was presented, estimated by rating and averaging the presence of various threats to validity. This percentage appears fairly small. However, quite a number of meta-analysts exclude studies on the basis of 'bad' quality, thus indirectly commenting on the adequacy of the included studies. This procedure is especially characteristic of work originating from the scientific community surrounding the Kuliks, their work making up about one fifth of the present sample. On the other hand several aspects relating to the quality of studies are recorded by meta-analysts.

The sampling method or assignment rule used in the primary studies was reported by 46% of the sample. In the analyses based on this classification, random sampling was usually interpreted as indicating higher quality. The quality of outcome measures, as represented by some index of the reliability of the instruments used, was coded in 63% of the cases, 46% reporting actual reliability coefficients while 17% contrast standardized tests with 'home-spun' measures. The issue of interaction effects or mediating variables controlled in the primary studies is recorded by 40% of the sample, thus indirectly touching the question of the quality of the research implementation. The adequacy of the statistical analysis in the primary studies, usually in the sense of the statistical power function of the test, was reported by 21%. Only 13% report the unit of analysis used in the studies. Thus, one known possible source of effect size variation was rarely examined.

Taken as a whole, only about 19% of the meta-analyses mention none of these issues of quality. Therefore, even though direct critiques of the primary studies are rare, at

least as many as reported by JACKSON (1980) concern themselves with similar problems.

6.2.4 *Data analysis*

As coding represents the measurement in meta-analysis, its quality is of the utmost importance for the reliability and validity of the study. One of the main problems encountered in coding is that of missing information. JACKSON (1980) reports that only one of his reviewers noted the missing data problem and tried to obtain the information elsewhere. In the present sample 67% mention the problem and 17% made an effort to gain the information from other sources. Having to exclude studies from analyses on this account is explicitly reported by 38% of the sample. This percentage does not include the meta-analyses which specified having sufficient data as part of their selection criteria, but only those that found after selecting studies that these did not report the information needed for coding specific categories. As a consequence of missing data, the sample sizes on which specific analyses are performed systematically decrease in numbers, thus in effect reducing their representativeness.

The reliability of the 'measurement' can be determined by analyzing the intercoder agreement. Despite the emphasis meta-analysts place on scientific standards, only 25% of the sample report some form of global intercoder agreement and 4% (two studies) present this information for more specific coding categories (one of these being the study by STOCK et al. (1982) which explicitly made a point of examining the variations in coding reliability across different categories). Specific strategies were adopted to reach intercoder agreement in 19% of the meta-analyses. Taken as a whole, these results were disappointing, as it had been expected that more interest would be shown in ensuring a reliable database.

Whereas JACKSON (1980) reports that about 80% of his sample were ambiguous on how they represented findings when analyzing them, 98% of the present sample indicate, in varying degrees of specificity, the index or formula utilized for computing the effect size. Whether this information was precise enough to allow replication would have to be examined by experts in meta-analytic methodology.

If the sample size magnitudes were listed per primary study, an attempted replication could clarify this question by direct comparisons of the computed effect sizes. However, only 35% report the effect size either for every study or larger subsamples of studies (8%). FRICKE and TREINIES (1985) also coded this aspect, mentioning that 22% present the effect sizes per study. In actual fact, it should be 24% as they missed coding one study (BANGERT-DROWNS, KULIK & KULIK, 1983) which, after double

checking, did contain the information. Ignoring the percentage with incomplete sample coverage, the two values (27% / 24%) correspond quite well.

In their study FRICKE and TREINIES (1985) interpret the presence of both sample size and effect size per study as allowing the meta-analytic computations to be replicated. In both their and the present study 15% fulfilled this requirement. No coding differences were identified in this case. In a sample of 22 meta-analyses, published mainly in 1983, which Fricke and Treinies analyzed after their manuscript had been handed to the publishers, 27% were replicable by this criterion, leading them to the conclusion that matters are improving. If, however, one takes the data as one sample, a procedure that seems justified seeing that the previous 67 contained several studies dating from 1983 (cf. Figure 2, p.68), this percentage decreases to about 18%, thus more or less what was found in the present study.

When readers are not given the exact effect sizes per study, their grasp of the range or variability of the findings would be enhanced if the authors were to present the information in the form of graphic frequency distributions or stem-and-leaf tables. This would at least allow the identification of outliers or other distribution irregularities. In 48% of the cases this was done either for the entire sample or larger subsamples of studies (13%). A total of 16 meta-analyses (33%) list neither the effect size per study nor stem-leaf or graphic summaries. The outcome variable involved in either kind of presentation is explicitly referred to by 60% of the sample, thus enabling readers to make even more concrete interpretations of the information.

Although JACKSON (1980) reports data relevant to these aspects, a comparison proved difficult as their meaning was not quite plain. On the one hand he states that 22% (8/36) do not report the findings of most of the reviewed studies and do not indicate how many or what percentage had each type of result, on the other that 50% (18/36) represent any of the findings with an indication of the direction and magnitude of the results, few for each study. If one interprets this as meaning that in about half of the reviews the reader is unable to form an impression on the distribution of the results, the advent of meta-analyses seems to have improved matters, 67% reporting data giving the overall picture.

Similarly, the number of authors passing on information on the significance of the findings seems to have increased: whereas JACKSON (1980) reports that only 11% (4/36) make clear distinctions between positive or negative and significant positive or negative results, 25% of the present sample indicate this either per primary study or through a vote count procedure (10%). FRICKE and TREINIES (1985) report that 13% of their sample used vote count procedures as integration method, which corresponds to the percentage identified in the present sample, not counting those listing the significance per study.

Effect sizes are the unit of analysis in meta-analyses. The actual analyses can, however, be based on either the effect size per study or effect size per study findings, the former indicating that the authors are aware of the problem of non-independence created by including more than one result per study in their analysis. The meta-analysts are not particularly explicit in reporting this information. While 40% were coded as using one effect size per study, 65% seemed to be using the study findings (17% of the codings were uncertain, in another 4% the information was not even adequate enough to allow tentative coding). The problem was aggravated to some extent by the confusing way in which the sample sizes were reported (cf. p.57), which could otherwise have helped clarify the question.

This ambiguity is underlined by the fact that although 67% indicate that the problem of non-independence was present and taken into account by one of a variety of procedures, 52% were coded as ignoring the problem (in either category 13% of the codings were uncertain). Another aspect of non-independence, specifically pointed out by methodologists, created by including several studies by the same author or research team in the sample, was ignored in 58% of the cases. As 35% do not publish a list of the articles (cf. p.57), the problem is bound to be even more prevalent. Not one of the meta-analyses even mentions the issue.

The extent to which these deficits invalidate the statistical analyses would have to be examined by experts. It came as a surprise, however, to find so much inexact reporting on such a widely discussed topic. One explanation for this state of affairs could be that as yet no completely acceptable way of dealing with the problem of non-independence has been found.

The average effect sizes computed either for the total sample or various subsamples of studies are a source of information not available in the traditional narrative reviews. In meta-analyses, however, this gives the reader another way of gaining insight into the findings of the primary studies, should these not have been presented individually or in summarized fashion. While 56% report the average effect size for the complete sample, 98% do so for subsamples. To be able to interpret these values adequately the reader should be presented with information concerning either the range, standard deviation or standard error. Although 90% were coded as mentioning any one of these measures of variability, this does not mean that every average was accompanied by one. The category was recorded as present, if any one of the averages found in the meta-analysis was reported together with some index of variability.

Even though JACKSON (1980) was not in a position to provide results on this aspect, the following findings indicate some of the imperfections of narrative reviews on a related subject: only one study (3%) discussed the full set of located studies, 17% clearly did not and in 80% there was not enough information to make a judgement. The analysis

by FRICKE and TREINIES (1985) contains coding categories similar to the ones employed in the present study. However, whereas they do not differentiate between the average effect size per sample or subsample, the present evaluation does not differentiate between standard deviation and standard error, additionally including range data in coding the category. The percentages are thus not directly comparable. The authors report that 91% of their sample presented average effect sizes, 45% the standard deviation and 37% the standard error. Using the data published in their coding table, it was established that 65% of their meta-analyses report either one or the other measure of variability, the difference in findings most probably being due to the much broader conceptualization of the category in the present study.

A similar disparity between the two studies exists in coding the type of integration procedure used. Whereas the present study employed a rather global and arbitrary differentiation of 'conventional' and 'modern' techniques, FRICKE and TREINIES (1985) distinguish between ten different procedures. The conventional analyses encompassed various techniques such as t-tests, ANOVA, regression analysis or vote counts. The modern techniques included combined probability tests and methods developed by Hedges or Hunter, Schmidt and Jackson to improve on identified statistical inadequacies. The use of conventional analyses predominates (75%), only 25% utilizing what was designated as modern techniques. The trend is similar to that found by FRICKE and TREINIES (1985). They report that 84% utilized the methods propagated by Glass, McGaw and Smith (i.e. conventional), a much smaller proportion using the 'modern' techniques (ranging from 0% to 24% depending on the specific method employed, a result not altered by the 22 additional studies they analyzed).

Although the data should fulfil certain requirements to allow the valid application of statistical procedures, only 8% report testing these assumptions. Another 27% point out limitations inherent in the techniques, thus warning readers not to over-interpret the results. As yet, there is no agreement among meta-analysts as to which of the available methods is to be preferred. Nonetheless, only 29% perform parallel analyses to ascertain whether similar results are obtained. Using the data recorded by FRICKE and TREINIES (1985), a percentage nearly identical to the one established for the present study was found: 30% employed more than one integration technique. A problem created by multiple statistical testing is the capitalization on chance. Most meta-analyses perform more than one statistical analysis on the same set of data, however, only 8% mention the fact that chance alone might be responsible for the significant effects found.

These multiple tests are frequently carried out in the search for variables which could possibly be influencing the magnitudes of effect sizes. Some methodologists maintain that the search for mediating variables should only be undertaken after establishing that the effect sizes are not homogeneous, otherwise this could lead to an over-interpretation of the data. However, only 23% used techniques developed to examine the heteroge-

neity of samples. Although one cannot directly compare what JACKSON (1980) examined to the present data, it might be of interest to note that 50% of his sample did not provide information for judging whether the reviewer interpreted the variations in findings in the light of sampling error. This situation does not seem to have changed. Nor is the impression altered by the findings FRICKE and TREINIES (1985) report: 9% of their 67 meta-analyses performed heterogeneity tests.

Nonetheless, most meta-analysts examine the variability of effect sizes by analyzing the mediating effects of diverse coded 'independent' variables. Analyzed most frequently are various categories of outcome variables (90%), followed by variations in treatment implementation and context or scope variables (both 81%). Subject characteristics were examined by 79% of the sample, 50% analyzed different aspects relating to the quality of the studies and 42% inspected the influence of more specific design features or aspects of the statistical analysis as possibly moderating effects. JACKSON (1980) reports that 19% (7/36) of his reviewers discussed or analyzed the relation of study features to findings. In descending order of frequency these were treatment or causal variables, subject characteristics, design or statistical analysis and contextual variables. Thus, the effort made to identify possible mediating influences has increased considerably.

All in all, it appears that meta-analysis has led to an increased awareness of the need for systematic presentation and analysis of the results of the primary research being integrated. However, there still seems to be much room left for further improvement, especially to achieve the high standards meta-analysts have set themselves. In particular, those issues identified as problematic should be taken note of and discussed in greater detail.

6.2.5 *Interpretation*

The potential usefulness of the meta-analyses is influenced by the way the presented data and results are interpreted. Fundamental to understanding the quantitative data, is knowing what the numerical values of effect sizes actually imply. Most meta-analysts make an attempt to clarify their meaning, only 25% do not. Predominantly the authors employ standard deviations or percentiles for this purpose (71%), 29% use Cohen's classification of small, medium and large effects, frequently as additional interpretational help. Only 10% try describing what effect sizes mean through behavioral indices. In most cases this was done by translating the effect size into the expected increase of scores on standardized tests. Expert judgements and the binomial effect size display were rarely utilized (only 2% in either case). This corresponds fairly well to the 3% FRICKE & TREINIES (1985) report as having presented the binomial effect size display.

Knowing how to interpret the effect sizes does not mean that one has also grasped their theoretical implications or how well they correspond to the prevalent theories and findings in the field. It was found that 83% compared their results to previous research, indicating whether they seemed to confirm or disprove old theories, hypotheses or impressions. This high percentage came as a surprise as reading the meta-analyses had created a different impression. The percentage does not reflect the considerable variation in depth and quality with which results are related to former work. A relatively large number are rather vague on the subject, alloting no more than a couple of sentences to the topic. Others placed their findings more solidly within the context of prior review or meta-analytic results, trying to integrate the new and the known or clearly casting doubt on previous interpretations.

Only 10% were recorded as attempting to formulate new theoretical aspects on the basis of their results. The coding was uncertain in four of the five meta-analyses involved. This is due to the fact that the reported information did not fit the category adequately. New theories or hypotheses were not really developed, rather they seemed to be theoretical modifications suggested by specific mediating variables identified in the analyses.

Whereas in the present study 15% were coded as mentioning neither of the categories, JACKSON (1980) found that 19% (7/36) confirmed or disproved old theory or induced and reported new theory. Thus, apparently matters have improved, but the ideal aimed for has as yet to be reached.

For practioners the formulation of implications for policy and practice is likely to be of greater importance. However, only 38% of the meta-analysts attempt to do so. That most of these did not turn out to be very concrete, came as no surprise. As the research itself covers specific sections of reality, making comprehensive recommendations is difficult and can only be tentative, which might deter authors from taking a more explicit stance. This impression is supported to some extent by the fact that 71% indicated factors mediating or limiting the conclusions for both theory and practice. Even so, compared to what JACKSON (1980) reports, the situation seems to have improved: 17% (6/36) stated recommendations for policy or practice, only four of the six studies mentioning conditions which might affect the impact.

On the question of formulating implications for primary research, however, the comparison of the two studies tends to favour the narrative reviews. While JACKSON (1980) found that 78% suggest desirable focuses or methods for future primary or secondary research, only 67% of the meta-analyses do so. It had not been anticipated that only 25% would mention implications for future meta-analyses or reviews. Although this percentage exceeds the 8% JACKSON (1980) found, one would have expected more relevant suggestions, considering that most meta-analysts hold methodology and

scientific standards in high regard. Furthermore, as this approach to integrating research is still fairly new, it seems probable that specific difficulties arise during the work that could be handled more effectively in future, if they were reported. The absence of a more critical evaluation of their own work also came as a surprise. Apparently the critical attitude was restricted to the more methodological articles rather than expressing it in the course of conducting and reporting an actual meta-analysis.

6.3 Do critiques or replication attempts affect evaluations?

The assessment of this question is restricted to impressions gained from critiques or replication attempts of the meta-analyses by JOHNSON, MARUYAMA, JOHNSON, NELSON and SKON (1981), KLAUER (1981), GLASS and SMITH (1979) and SMITH and GLASS (1980), an article by SLAVIN (1984) critically examining several meta-analyses and reanalyses made by FRICKE and TREINIES (1985) concerning the question of heterogeneity (cf. Appendix 1, indented sections). The impressions are thus not representative, but do show general trends.

The theoretical and ideological controversies present in the field, a fact often glossed over in the actual meta-analysis, are shown up by these articles. Readers are warned not to accept conclusions uncritically because certain aspects have been ignored. They are cautioned that issues are not as straightforward as implied. The lack of specificity in defining the constructs is another matter frequently referred to. The diversity, often not explicitly described by meta-analysts, is made evident by pointing out critical distinctions or theoretical implications ignored by the broader conceptualizations in the meta-analyses. These can lead to difficulties in the interpretation of results. Furthermore, they cast doubt on the representativeness of the research being integrated because evidently not all of the research pertaining to the broad definition was included in the sample, creating the impression that the inclusion or exclusion of studies was to some extent haphazard. Also indicated are problems concerning the statistical analyses, demonstrating that other techniques could lead to different conclusions or that limiting their applicability seems warranted.

In general, these papers sensitize readers to problems, but they neither solved any of the controversies nor changed the basic impression gained in evaluating a meta-analysis by the method outlined from section 5.1 onward. They do, however, create an awareness of the difficulties involved and present more specific details and expert assessments to points only vaguely identified as inadequate during the evaluation process.

6.4 *Is the quality of meta-analyses improving?*

To answer this question on the basis of the present sample is problematic for several reasons: some pertain to the size and representativeness of the sample of meta-analyses, others to the nature of the categories employed in coding the meta-analyses. Nonetheless, after discussing these difficulties, a description of possible trends will be attempted.

6.4.1 *Problematic issues: sample size and representativeness*

In contrast to the evaluation presented in section 6.2, all the meta-analyses, regardless of whether they were considered dependent or independent, were examined to determine possible trends, since all of the dependent studies were published in different years (cf. p.49 & Figure 1). This sample of 55 meta-analyses is really too small to allow an accurate identification of trends for the 10 year period ranging from 1979 to 1988. Assuming a uniform distribution of the studies across the years, would mean that an average of only 5.5 meta-analyses would be obtained for each year. This is a very small basis for a trend analysis, quite apart from the fact that the assumption of a uniform distribution is unrealistic.

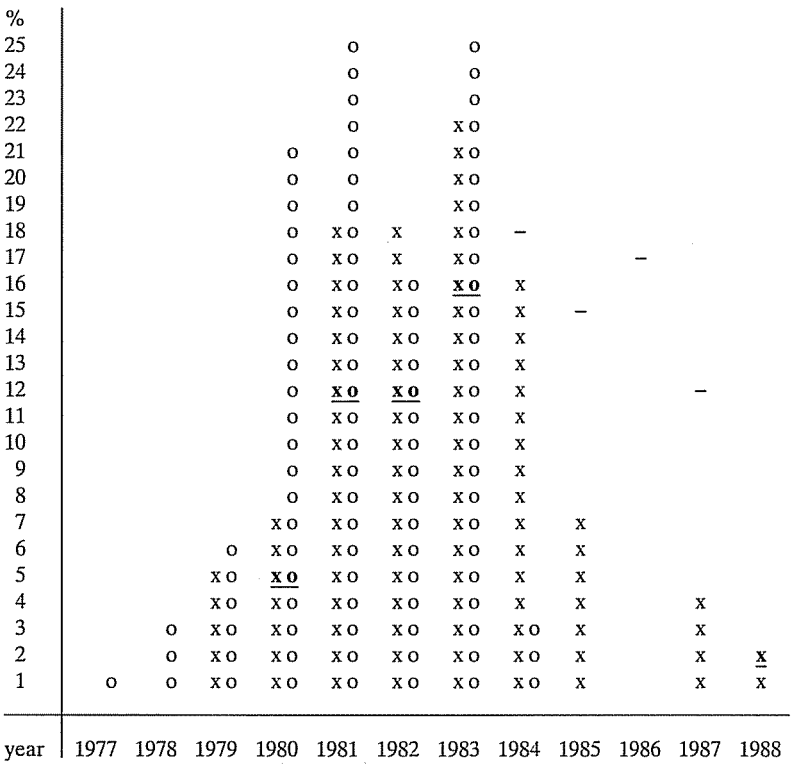
Because the meta-analytic approach was only developed in 1976, it is reasonable to presume that an increasing number of studies using the approach will be published as more scientists become acquainted with it. In the present sample the distribution of meta-analyses across the years is roughly bell-shaped, ranging from 0 in 1986 to 12 in 1983 (cf. Figure 1, p.50). To ensure that the database per time unit is at least minimally adequate, it was necessary to combine the meta-analyses published in 1979/1980 and 1985 to 1988, leaving a total of only 6 time intervals to analyze. The small sample size per unit of time, a minimum of 7 to a maximum of 12 meta-analyses, the lack of independence of 7 of the meta-analyses and the relatively low number of time intervals have to be taken into account when interpreting the results (cf. Appendix 3 for the coding frequencies and percentages per year).

Performing a trend analysis implies that the sample is representative of the kind of meta-analyses published during that period. Whether this holds true for the present sample is questionable. Firstly, representativeness was not a guiding principle in assembling the sample. The educational variables examined by the meta-analyses were of primary interest (cf. section 6.1.1), thus leading to a selective sample. Secondly, as expected, the percentage of meta-analyses per year increases until 1983. After that, however, instead of leveling off, having reached a sort of average number of published meta-analyses per year, it decreases again (cf. Figures 1 & 2). This does not imply that the ap-

proach is being used less often. Rather, it appears to be the result of the sampling procedure, as the following will show.

Comparing the percentage distribution of the meta-analyses examined by FRICKE & TREINIES (1985) with that of the present sample, Figure 2 shows that up to 1983 there are similar increases. As their book was published in 1985 the abrupt decrease in the number of the meta-analyses examined for 1984 seems logical: their analysis must have

Figure 2. Percentage of evaluated studies per year of publication
(x=present study , N=55; o=Fricke & Treinies (1985), N=68)



The total number of studies evaluated by Fricke & Treinies (1985) is 68 instead of 67 because in one case they counted two articles published by the same author in different years as one (Klauer, 1981, 1984). The horizontal lines indicate the percentage distribution of the 149 references to meta-analyses obtained through the PSYCINFO and PSYNDEX search.

documented in the databases of the retrieval systems.

These arguments do not apply in the same way to the present sample. The search for literature was conducted in 1988. Thus, a similar decrease should only have occurred in 1988. By examining the percentage distribution of the 149 references to meta-analyses obtained through the search in PSYCINFO and PSYNDEX, which was conducted to find studies relevant to the theoretical network of educational variables (cf. p.48/49), it became evident that the decrease in the number of meta-analyses after 1983 is definitely an artifact of the sampling procedure (meta-analyses not pertaining specifically to the theoretical network nor promising essential information on methodological issues were not ordered from the library for inspection). The distribution of the 149 references shows the expected increases until 1984, then levels off to decrease again in 1987, with a sharp drop in 1988, the year in which the search was conducted (cf. Figure 2, horizontal lines).

As a result of these observations it seems plausible to assume that until 1984 the present study contains a fairly representative sample of meta-analyses, an assumption which is underlined by the similarities between the present study and the results obtained by FRICKE & TREINIES (1985), described in section 6.2. However, for the years 1985 to 1988, which had to be combined into one category for the purpose of identifying possible trends, it is doubtful whether the sample fulfills this requirement.

6.4.2 Problematic issues: coding procedure

The coding scheme, consisting of 80 categories, was devised to document how various critical aspects of the meta-analytic process are typically handled. The categories have a predominantly descriptive function. They do not necessarily reflect the quality of a meta-analysis as such and were not intended to do so.

The aspects recorded as present in an article cannot be added up to give an overall mark expressing the quality of a meta-analysis. To make such a number interpretable, it would have been necessary to classify each item as either positive, negative or neutral, an undertaking that is not without hazards for the following reason: an item that individually appears to have a specific value can obtain a completely different meaning when seen in combination with others or within the context of the meta-analysis as a whole.

For example, ignoring the diversity of the theoretical concepts of the primary studies analyzed by the meta-analysis would seem to indicate a negative quality. Yet, if the aim of the meta-analysis is to study such a broad, diffuse concept, ignoring the differences is in keeping with the intentions and cannot be regarded as negative. An apparently neutral aspect such as recording and analyzing gender influences, or neglecting to do so, can

obtain a decidedly different quality if these are known to have an essential influence in the subject area being integrated. Seemingly positive items such as indicating the formula used to calculate the effect sizes or testing whether the assumptions underlying the statistical analysis have been fulfilled, acquire a negative meaning if the formula are applied incorrectly or the data do not meet the necessary requirements.

In effect, this means that the quality of a meta-analysis has to be determined individually, weighting the various aspects according to the overall context. What the coding scheme covers are those issues that meta-analysts consider essential to ensure that the research integration can potentially fulfil rigorous scientific standards and have high quality. Neglecting to report specific critical issues cannot necessarily be equated with neglecting to consider them in conducting the work (compare MEINEFELD, 1985, p.300). Thus, the coded data indicate indirectly to what extent the meta-analyses fulfil the standards the meta-analysts have set themselves. However, adequate coverage allows readers to be intelligent consumers rather than mere recipients of the information. Reporting these aspects puts them in a position to critically analyze a meta-analysis and assess its quality.

In attempting to answer the question of whether the quality of meta-analyses is improving on the basis of the data available in the present study, these issues have to be kept in mind. The data primarily indicate whether sufficient information is reported in a meta-analysis to allow a critical examination of its quality. However, if authors show a critical awareness of these problematic aspects of the meta-analytic approach it is likely that they have taken some measures to ensure that their work is of adequate quality. Presenting readers with enough information enables them to decide independently whether the results and interpretations are reliable and valid enough for their own purposes.

So, what can basically be established when examining the present data for possible trends is whether the quality of reporting, which indirectly reflects the potential quality of a meta-analysis rather than determines its quality as such, is improving. These trends will be discussed in the following section.

6.4.3 Possible trends in the reporting quality of meta-analyses

The percentage of meta-analyses recorded per item category in each time interval was determined as well as the average percentage of the group of 55 meta-analyses as a whole, i.e. the average percentage for the total period of time (cf. Appendix 3). To analyze whether possible trends exist the percentage coverage of an item in each time interval was compared to that calculated for the total group. If trends exist, this should show

Table 2. The number of the 80 item categories per time interval with percentages below or above the respective average percentage calculated for the 55 meta-analyses as a whole.

interval	number of items ‹ = average percentage		number of items › average percentage	
1979/80	47	(5)	33	(2)
1981	38	(2)	42	(5)
1982	43	(4)	37	(3)
1983	34	(1)	46	(6)
1984	52	(6)	28	(1)
1985-88	39	(3)	41	(4)

The numbers in parenthesis indicate the rank order.

itself in either systematic increases or decreases across the years. However, the percentages per time interval and item appear to fluctuate rather haphazardly around the average percentage calculated for the total period.

That this is so, is roughly exemplified in Table 2. It shows in how many of the 80 item categories the percentage of recordings per time interval was above or below the average determined for the item. As can be seen there is no regular upward or downward trend across the years. In general, if an item category was recorded as present in a certain percentage of meta-analyses, this tended to remain fairly stable over the years with seemingly chance or unpredictable fluctuations about the average. There are, however, some interesting exceptions.

The percentage of meta-analyses publishing lists of the primary studies being integrated appears to be increasing steadily (from 14% in 1979/80 to 71% in 1985-88). At the same time the percentage of meta-analyses for which these articles were available to readers at request only decreased (from 86% in 1979/80 to 0% in 1985-88). In both cases the percentages established for 1981 seem to anticipate these trends (cf. Table 3). Both trends seem to indicate an improvement in the reporting quality because readers are given information enabling them to assess the scope and representativeness of the meta-analyses more specifically.

These improvements could have occurred because meta-analysts have come to appreciate the importance of imparting this information. It seems likely, however, that purely technical publishing considerations play a decisive role in this process. This is suggested by the fact that the medians of the number of primary studies integrated by the meta-analyses of a specific time interval follow an analogous trend (cf. Table 3). As the number of integrated studies decreases it is easier to publish a list of the articles and con-

Table 3. Medians of the number of primary studies integrated by the meta-analyses and the percentage of recordings per item categories revealing possible trends

	1979/80	1981	1982	1983	1984	1985-88	Tot*
median of primary studies integrated by meta-analyses	72	36	53	54	24	14	43
articles listed	14	80	40	67	67	71	58
articles available at request	86	10	40	17	22	–	27
effect sizes listed per study	–	20	30	25	56	71	33
stem-leaf/graph summary	71	40	60	25	44	29	44
conventional stat.analysis	86	90	60	67	67	57	71
'modern' stat. analysis	–	10	20	58	22	29	25
heterogeneity tested	–	10	30	25	22	29	20
assumptions tested	–	–	10	17	11	14	9
specific hypotheses	–	10	10	25	33	14	16
quality aspects	57	70	40	42	44	43	49
contextual/scope variables	100	80	80	75	67	57	76
reliability of outcome	71	70	60	58	33	43	56
interaction/control var.	43	40	50	33	33	14	36
Construct definitions:							
indirectly deducible	86	40	70	58	56	43	58
specifically formulated	57	90	60	75	89	71	75
missing data problem noted	86	80	60	67	56	43	65

* Tot. refers to the values calculated for the total of 55 meta-analyses. The respective sample sizes for the six time intervals were 7, 10, 10, 12, 9 and 7.

sequently less necessary to have to provide them at request only.

Similar considerations could explain the upward trend in the percentage of meta-analyses that list the magnitudes of effect sizes per primary study and the corresponding, though not so pronounced, downward trend in the percentage of stem-leaf or graphical representations of effect sizes (cf. Table 3). Both could plausibly be accounted for by the size of the sample being integrated. As the number of primary studies decreases it would be more feasible to provide readers with individual results rather than only presenting graphical summaries.

Unfortunately, it is impossible to establish with certainty whether the trend towards integrating smaller samples of primary studies is an artifact of the present sample of meta-analyses or whether it can possibly be generalized. Only 92 abstracts of the 149 references to meta-analyses obtained through the literature search contain information

about sample sizes. The years 1980, 1987 and 1988 were excluded from the investigation as they respectively contained only 5, 3 and 1 meta-analysis to estimate the median number of primary studies. For the remaining total of 83 meta-analyses a similar trend in the median sample size could be observed: declining from 74 primary studies in 1981 to respectively 53, 54, 39, 32 and finally 22 in 1986. Because of this and judging by the amount of criticism directed at the practice of including very diverse primary studies, regardless of their quality, in a meta-analysis, it would not be unreasonable to presume that such a trend exists.

Other trends seem to reflect the methodological developments in the meta-analytic approach. Across the years the percentage of 'conventional' statistical analyses seems to be decreasing, though maintaining a relatively high level (86% in 1979/80 to 57% in 1985-88, cf. Table 3). In comparison, as expected, the 'modern' analysis techniques make their appearance in 1981 (10%) with increasing tendencies, though remaining at a fairly low level (under 30%) except for 1983 (58%). The same applies to the upward trend in testing the heterogeneity of effect sizes, after the development and introduction of the appropriate techniques. Like the modern approaches to analyzing the data, however, its use remains at fairly low levels.

Whether the analogous trend shown in the percentage of meta-analyses testing the assumptions on which their statistical analyses are based (cf. Table 3), can also be explained by methodological developments is doubtful. This tendency seems to reflect an increased (though generally low) awareness of the need to test assumptions to guarantee valid analyses, perhaps sparked off by the critical discussions of methodological aspects.

A similar awareness might account for the upward trend discernible in the percentage of meta-analyses formulating specific hypotheses as aims, though the percentage that do so remains fairly low (cf. Table 3). Formulating specific hypotheses is mentioned as a possible way of counteracting some of the problems arising due to multiple testing with the resultant capitalization on chance (cf. p.27).

An interesting trend is the apparent gradual decrease in the percentage of meta-analyses examining quality aspects of primary studies as possible influences on effect size variability (cf. Table 3). This could indicate that meta-analysts are less preoccupied with questions of quality in their analyses because they have started to take care of the problem while selecting the primary studies. Perhaps they are heeding the arguments, voiced soon after the introduction of the meta-analytic approach, that the meta-analysis can only be as good as the data and studies included. Whether this is so cannot be determined with certainty as the relevant data concerning the selection criteria were not coded for the present sample.

A similar downward trend is visible in the percentage of meta-analyses examining contextual or scope variables as possible sources of between study effect size variability (cf. Table 3). Why this should be the case is incomprehensible and seems to be totally inexplicable.

However, possibly these two downward trends are somehow linked to two analogous ones recognizable in the coding of study features: reliability of outcome measures and interaction or control variables (cf. Table 3). The reliability of outcome measures can be considered an aspect of primary study quality. Similarly, interaction and control variables can be regarded as forming part of a primary study's contextual variables. So, if there is a downward trend in either one of these, a corresponding trend should exist in either one of the others.

Furthermore, presenting less information on the reliability of outcome measures and the interaction or control variables of primary studies in a meta-analysis seems to reflect negatively on its reporting quality. It restricts the extent to which readers can form an independent opinion of the trustworthiness and scope of the primary studies and consequently that of the meta-analyses.

A trend that appears to reflect positively on the reporting quality of meta-analyses is the decreasing percentage of meta-analyses in which construct definitions had to be deduced indirectly, even though it remains over 40% (cf. Table 3). However, the nature of the coding scheme employed in the present study does not lend itself to determining whether this is really the case. Another possible interpretation of the declining percentages is that no information on the construct definitions was discernible in the text at all. This possibility cannot be excluded on the basis of the present data. The fact that a corresponding increasing pattern is not mirrored in the percentages of meta-analyses coded as specifically formulating the construct definitions rather seems to underline this possibility.

A fairly clear downward trend appears to exist in the percentage of meta-analyses noting that the primary studies do not report the data necessary for coding relevant categories or calculating effect sizes (cf. Table 3). This would seem to reflect qualitative aspects of the primary studies rather than the reporting quality of meta-analyses as such. Interpreted in this way, it could mean that the reporting quality of primary studies might be improving. This interpretation is fairly unlikely though, since it would mean that the meta-analyses in each time interval are integrating primary studies that have correspondingly been published more recently. This seems implausible, quite apart from the fact that the aspect was not coded in the present sample. Other possible interpretations could be that less detailed coding schemes might have been used in the meta-analyses so that missing information in primary studies would not be as noticeable or that studies not containing the necessary information were excluded from meta-analyses at the selection

stage. Neither possibility can be answered on the basis of the present data.

In general, even though these 15 trends appear to be relatively clear, their existence and possible meaning has to be seen in the light of the limitations mentioned in sections 6.4.1 and 6.4.2. Apart from this, there are 65 other item categories for which no such clear trends could be identified. It seems reasonable, therefore, to assume, as indicated at the beginning of this section, that no extraordinary changes in the quality of reporting meta-analyses have occurred during the past ten years, apart from slight shifts in emphasis, probably as a result of methodological developments. In addition, this means that the evaluation of the data presented in section 6.2 is not essentially affected and does not have to be reinterpreted in the light of possible developments in the meta-analytic approach occurring through the years. It therefore seems highly unlikely that the actual quality of meta-analyses has been improving.

6.5 Resumé

If one takes the results presented by JACKSON (1980) and WAXMAN and WALBERG (1982) as base-line for the quality of the typical narrative review, then the meta-analytic approach seems to have brought some improvement in most of the methodological aspects concerned. If, however, one takes the standards formulated for the approach by the meta-analysts themselves as base-line, one has to conclude, as FRICKE and TREINIES (1985) do, that the objectivity aimed for has, as yet, not been attained (also cf. ABRAMI et al., 1988). They restrict this evaluation to the statistical part of meta-analyses, believing that aspects such as problem formulation, sampling, assessment and interpretation of the findings can only be evaluated by experts in the field. Similarly, SLAVIN (1984) believes that it is difficult to evaluate meta-analyses without understanding the underlying studies. This point of view is not shared entirely.

The present evaluation could show, without having to take recourse to expert knowledge, that these issues are often not handled as efficiently as they could be. Non-experts can roughly assess a meta-analysis by sticking to the evaluation scheme outlined above. Utilizing these evaluation criteria they might not be able to pinpoint the exact nature of the deficits identified. However, they will be able to determine whether or not confidence in the results or conclusions seems warranted. Obviously, this procedure cannot be equated to an expert evaluation of a meta-analysis which is more than ticking off various criteria as being present or not. This is also pointed out by DRINKMANN (1990) in reporting the evaluation study conducted by SACKS, BERRIER, REITMAN, ANCONA-BERK and CHALMERS (1987) who found that of the 23 evaluation criteria they used only an average of 7,7 were taken into account by the 86 meta-analyses analyzed.

Some of the identified deficits could be resolved quite easily by more detailed reporting. As JACKSON (1980) suggested, if less than 40 studies are reviewed, it should be possible to present a single-page table indicating a few of the characteristics of all the primary studies as well as their findings. Strangely enough, he does not follow his own recommendations, neither publishing a list of the 36 reviews examined nor a table of his findings. In the present study 23 meta-analyses integrate 40 studies or less, however, only 13 of these publish details of findings per study.

In other respects it will prove more difficult to improve matters, a case in point being the interpretation of results. Most of the meta-analyses are exploratory in character and the theoretical framework is often rather vague. Both facts are not conducive to having results more firmly embedded within a complex theoretical background. If one accepts the characterization of meta-analysis given by GLASS (1983) as evaluative and atheoretical, serving to assess rather than explain, not intending to illuminate phenomena in an explanatory or analytic way, then just enumerating the findings might seem adequate. However, to synthesize and advance knowledge more is needed than summarizing the state of knowledge in the area (STRIKE & POSNER, 1983).

Part of the difficulty might stem from the nature of the research itself. If what TEDESCHI, GAES, RIORDAN and QUIGLEY-FERNANDEZ (1981) report for research in social psychology is also true for educational research (which seems possible as indicated by a limited analysis conducted in a study by HAGER (1985)), namely, that about half of the current research does not test available theory, then one cannot expect an integration to relate more specifically to theory. They found that most of the studies are intuitive demonstrations of a single idea which lead nowhere in terms of the accumulation of scientific knowledge, but also state that the situation was beginning to change. A more pessimistic view is taken by MEEHL (1978) who feels that possibly there will never be a really impressive theory in personality or social psychology, most so-called theories in the soft areas of psychology being scientifically unimpressive and technologically worthless.

Seen in this light meta-analysts perhaps cannot be expected to relate the findings more explicitly to theory. It does, however, not relieve them of the necessity of at least reporting as exactly as possible what the theoretical concepts being integrated encompass. This would enable readers to decide for themselves whether the findings could be applied to help solve problems or answer questions at hand. Furthermore, since the advent of the meta-analytic approach there has been no want of suggestions for improvements (e.g. KENDALL & MARUYAMA, 1985; KIESLER, 1985; STRUBE, GARDNER & HARTMANN, 1985). However, relatively few of these seem to have been actually adopted in practice. Some of those which appear especially useful for enhancing the quality of meta-analyses will be discussed in the following section.

7 Improving the quality and utility of meta-analyses

After analyzing eight meta-analyses to determine the degree to which the approach contributes to research integration, SLAVIN (1984) comes to the drastic conclusion that the way a typical meta-analysis is conducted and described is a significant step backward in the art of research synthesis. Although this harsh conclusion is not shared, one has to admit that meta-analysts are far from fulfilling the ambitious aims and functions they had conceived for themselves (cf. section 2.1.) and would be well-advised to follow some of the proposals offered for their enhancement.

The need for improvement is also emphasized by ABRAMI et al. (1988) after a detailed evaluation of six meta-analyses on student rating validity. The aim of their study was to examine how well quantitative reviewing methodology has been implemented. They found that the conclusions of the six meta-analyses were not similar. The differences were not limited to technical details and occurred at each step of the reviewing process (formulation of inclusion criteria, location of studies, coding, quantification of results and aspects of data analysis). There were also differences in conception and execution. In their view, the differences do not reflect shortcomings of the methodology but rather show that there are problems in the implementation of the methodology. Consequently, they suggest various ways of resolving these by improving the specification of inclusion criteria, coding of study features (for a practical example cf. ABRAMI, d'APOLLONIA, & COHEN, 1990) and data analysis.

In their enlightening article CORDRAY and ORWIN (1983) suggest ways of exploiting the overlap between the three levels of research (primary evaluation, secondary analysis and quantitative synthesis) to improve the utility and quality of evaluation efforts. In essence, they recommend making more extensive use of the information available or attainable at each of the levels. Attending to more of the methodological characteristics in primary studies would allow quantitative syntheses to accumulate results beyond the substantive conclusions. This in turn would provide a normative or actuarial database on the relative merits of designs and procedures, on flaws and rival hypotheses or on conditions likely to be encountered in future research, all of which would be a valuable resource for planning new research. This sentiment is shared by KULIK (1984) who feels that accumulating and comparing meta-analytic findings from different areas would lead to a better understanding of factors influencing the outcome of research (also cf. PILLEMER & LIGHT, 1980). Secondary analysis of primary studies or at least pointing out the flaws limiting their validity and utility could help to improve the general quality and informational content of the meta-analytic database. Similarly, critiques of quantitative syntheses could highlight controversial issues leading to healthy controversy (FISKE, 1983), just as reanalyses, using competing techniques or redundant

data-sets for cross-validation, could help in assessing the validity or robustness of meta-analytic results (also cf. HEDGES, 1986).

In the same vein DRINKMANN (1990) advocates taking a multi-method approach in meta-analyses. On the one hand, depending on the technique chosen, different results could be obtained, on the other the technique chosen might result in possibly systematic selection effects. The extent to which these might influence the results cannot be predicted. Performing parallel analyses would lead to data useful for analyzing these aspects, add to knowledge about the appropriateness and advisability of choosing specific techniques as well as allowing one to estimate and perhaps control the extent to which selection effects determine results.

These useful recommendations remain largely within the strictly quantitative integration strategy followed by meta-analysts. However, as HEDGES (1986) puts it, the most persuasive meta-analysis is likely to be one that combines the strengths of qualitative reviews with serious quantitative methodology (a belief shared by several methodologists, cf. end of section 3.3.3).

7.1 Combining qualitative and quantitative reviewing approaches

From the readers' point of view meta-analyses would benefit decidedly, if more of the qualitative way of reviewing were included in the quantitative approach. This concerns primarily the theoretical part of the meta-analysis. More explicit descriptions of the theories and constructs, the research techniques used in the area and the problems or controversial issues involved, would put the consumer in a better position to understand the rather abstract quantitative presentation and analyses which of necessity have to work with terse, scanty or reduced informational units, perhaps intelligible only to those familiar with the studies or field of research.

Besides this, presenting details narratively would diminish the slightly overwhelming effect created by the statistical data and allow consumers access to the findings from a less abstracted level. Through detailed qualitative and conceptual arguments they could gain a clearer insight into the nature of the theory and research being integrated than they can now with mainly the coded information to go by. This would reduce their having to rely predominantly on the interpretations given by the meta-analysts, rather than being able to reconstruct these by themselves.

The qualitative approach should be used to supplement the quantitative methods (CHELIMSKY & MORRA, 1984; HEDGES, 1986). This is necessary not only to enhance the understanding of the substantive domain but also because not all of the data presented in primary studies is amenable to being coded or fitted into the current frame-

work of the meta-analytic approach. One consequence is the loss of information, another that most meta-analyses exclude studies using other than comparative methods (CORDRAY & ORWIN, 1983), thus in effect basing the integration on a very specific selection of research evidence. Dealing with the data narratively is one way of coping with the problem.

On the other hand, qualitative methodology can be utilized to analyze descriptive features and findings as GUSKIN (1984) suggests. That this can be done profitably is demonstrated in the meta-analysis conducted by MOORE and READENCE (1984). The fact that meta-analyses have difficulty in capturing the contextual richness and subjective reports of study outcomes induced them to apply a form of content analysis to the researchers' discussions and conclusions. They could show that the aspects disclosed in this way can be employed to explain or mediate the results of the meta-analysis and help to develop new lines of research. In general, however, meta-analysts largely tend to neglect the qualitative data contained in the studies and to take little note of the possibilities qualitative approaches have to offer, although these could enrich meta-analyses in a variety of ways.

7.2 Taking communication quality into account

A perspective worthy of attention is central to the work of NOBLIT and HARE (1988). These scientists, working within the qualitative research tradition, perhaps triggered by the meta-analytic approach and the resulting discussion of research integration in general, realized that if they wish to communicate what interpretive research reveals to policy makers, concerned public or scholars and to reflect on their collective craft, they would have to find ways of synthesizing qualitative research. To this end they developed their approach called meta-ethnography, which is primarily concerned with understanding the sense of things. According to NOBLIT and HARE (1988) the translation theory of social explanation is the unique aspect of meta-ethnography. Communicating knowledge involves translating symbol systems between two or more parties. Because of this emphasis they concern themselves with the audience the synthesis is intended for. To be effectively communicated it should be rendered in a language and form appropriate to the audience. The worth of a synthesis is seen in its comprehensibility to some consumer.

It is this perspective that meta-analysts should take to heart. Although making research evidence available to consumers in an adequate way is formulated as one of the general aims of the approach (cf. section 2.1), judging by the typical format of the meta-analytic reports, the characteristics of the recipients of the information seem unclear. Perhaps the wish to reach the lay public as well as scientists, whether familiar or not with the domain being integrated, with the same article is part of the problem. The result is

that the needs of neither are appropriately satisfied. More attention to the aims of the meta-analysis, even if they are exploratory, might improve the situation, directing more explicitly what to analyze and report: integrating research evidence to achieve a normative basis for planning future studies in the field would need other details than necessary if the object were to inform practitioners about what research findings could possibly be transferred into practice. In either case the practical relevance of the meta-analysis would be improved.

These issues are discussed more explicitly in circles concerned specifically with the problem of knowledge accumulation and dissemination and appear far from resolved. As, however, the general presentation of meta-analyses could benefit from the reflections made, the following section will digress briefly to summarize some of the aspects considered important for knowledge synthesis and use in general.

7.3 Knowledge synthesis and practical relevance

Of special interest in the discourses on knowledge synthesis is the importance attached to the user perspective. This is evident in the definitions of knowledge synthesis, in the guidelines and criteria developed for 'good' syntheses as well as in the characteristics identified as determining the usefulness of knowledge.

There are two broad definitions of knowledge synthesis that represent two different perspectives: the first refers to the process of doing a synthesis, the second focuses on synthesis as a way in which knowledge is used. In close agreement with the aims of meta-analyses, synthesis is defined as the process of accumulating knowledge relevant to a given topic, showing the interrelationships among pieces of knowledge, moving from research reports to a consolidation and integration of findings, resulting in a product useful to various groups of practitioners (KLEIN, 1989; WARD & REED, 1983). These products include encyclopedic articles, research review reports or books that review specific areas of knowledge, all deriving their summaries, conclusions or implications from the work of others (KLEIN, 1989). However, emphasizing the users' perspective, synthesis is also seen as a process in the mind of the practitioner, policy maker or other knowledge user for utilizing knowledge to make sense of a situation or problem and to decide upon an appropriate course of action (WARD & REED, 1983).

Similarly, STRIKE and POSNER (1983) define synthesis as a multifaceted activity, unifying intellectual parts into a coherent whole. Based on this view, they develop a list of 15 types of intellectual enterprises that could be considered as cases of synthesis. They believe, however, that those involving a higher degree of conceptual innovation, generating a unifying conceptual framework, lie at the centre of the concept. Furthermore, in their conception, synthesis and research activities are closely linked (compare

CORDRAY & ORWIN, 1983), an integral part of the process of empirical investigation being the understanding and reflection of results which in turn directly influences the conceptualization of new studies.

Most of the reviews produced by the meta-analytic approach would fit into their category of quasi-synthesis. In their view, quasi-syntheses impose some form or order on ideas without generating an integrating perspective. They neither create new concepts nor modify, transform or reorganize current ones. Rather, they weigh the bulk of evidence from diverse sources, develop applications of an existing idea or assemble information in useful ways, yet falling short of intellectual wholes, i.e. only performing some of the roles of synthesis.

According to WARD and REED (1983) little is to be gained by labeling one document as synthesis and another as something else, because being identified as synthesis does not reflect on its quality or usefulness. Rather than trying to identify special classes of synthesis products, the crucial issue should be the relationship between the knowledge structure and the intended use of the knowledge, taking into account the knowledge needs, skills and settings of potential users. That the importance of these issues is also stressed by STRIKE and POSNER (1983) is expressed more explicitly in their discussion of what constitutes a good synthesis. On the one hand this question concerns the standards by which to judge the intellectual quality of a synthesis, on the other the standards for determining its usefulness.

In general, an intellectually sound synthesis will clarify and resolve the inconsistencies prevalent in the material being integrated, will result in a progressive problem shift and satisfy the formal criteria for good theory: consistency, parsimony, elegance and fruitfulness. In considering the aspect of usefulness, STRIKE and POSNER (1983) suggest that one has to qualify the question by adding 'useful for what?', thereby reintroducing the subject of the characteristics of the intended audience mentioned in the previous section. Although they do not believe that syntheses for researchers and practitioners should differ fundamentally, they suggest that these recipients indicate two contexts of use that should be examined.

From the practitioners point of view useful syntheses will be those that answer the questions being asked and present information relevant to fulfilling the consumers' objectives. This implies that useful knowledge provides recommendations that can be acted upon. However, it also assumes that practioners adequately understand the situation and are asking the appropriate questions. According to STRIKE and POSNER (1983) this tends to restrict the role of researchers to the more technical aspects of practice rather than also letting them provide theoretical information useful to understanding and deciding what might be worth doing. Moreover, this point of view makes practitioners mere recipients of research results rather than persons capable of understanding

the theoretical basis of the recommendations. To researchers a synthesis is useful if it does more than summarize research evidence in a given area. It should judge the current state of research and suggest future directions, helping to maintain a coherent research effort. They conclude, however, that closer cooperation between researchers and practioners would benefit research in general (also cf. GOOD 1983b).

The user perspective is also clearly included in the checklist of questions developed by KLEIN (1989) to guide the process of knowledge synthesis. They concern four characteristics of synthesis products: intrinsic qualities, spinoff 'unintended' effects, user-dependent qualities and development options.

Some qualities of syntheses are considered *intrinsic* because they do not depend on how the product interacts with specific users. As these questions concern highly specialized aspects, different experts might be necessary to guarantee them: the quality of the synthesis content, the technical or methodological quality, social fairness and communication quality. The latter, however, introduces the intended readers or audiences as factors needed to evaluate the quality of the document, aspects considered in greater detail under the heading of *user-dependent qualities*.

These relate to the desirability, utility or practicality and effectiveness of the synthesis. The question of desirability concerns issues such as whether there is a current demand for a synthesis on the specific topic and whether it could provide additional knowledge. The aspect of utility or practicality refers to matters such as the appropriateness for the intended user, readability, user-friendliness in the sense of being self-contained, accessible and presenting a sufficient amount of interpretation. Effectiveness relates more directly to how synthesis products affect users: Do they understand and learn from what they have read? Do they apply the knowledge? Does it change their behaviour?

The questions regarding the *spinoff benefits* concern more general effects: identifying gaps in the research, creating new insights, providing recommendations for improving research and synthesis efforts and the general impact as shown by references to or citations of the work in other accessible documents. The considerations KLEIN (1989) classifies under *development options* are intended to stimulate thought and discussion on what constitutes an effective knowledge synthesis process, a topic about which little is known.

Comparing these guidelines with the aims and functions meta-analysts have formulated for their work and what the typical meta-analytic reports are like, brings to light a few remarkable aspects. Whereas the aims and functions (cf. section 2.1.) seem to stress similar points, the actual meta-analyses seem to focus primarily on the issues KLEIN (1989) discusses under the heading of technical quality (i.e. the methodological adequacy); neglecting or only touching in a cursory way upon most of the other aspects con-

sidered necessary to obtain a knowledge synthesis of quality. One wonders, whether meta-analysts are aware of the fact that they are actually only performing a fraction of what it takes to synthesize knowledge. They seem to concentrate so much on methodology as to overlook that quantification, coding and statistical analysis alone do not provide the sort of comprehensive, organized information needed to benefit understanding and facilitate knowledge interpretation activities.

Although the strictly quantitative approach may enhance confidence in the conclusions reached, this alone does not make them useful to scientists and practitioners in the sense of presenting them with an adequate basis to work from. As CORDRAY and ORWIN (1983) point out, a quantitative synthesis cannot rely on the application of statistical procedures to yield meaningful results, a better balance between extracting statistics and interpretation needs to be achieved. To this purpose meta-analysts should perhaps take up some of the considerations discussed above and heed the characteristics WARD and REED (1983) tentatively identify as determining the usefulness of knowledge, i.e. its structure. Of the three major categories of characteristics (content, conceptual structure and physical structure) the following variables used to describe the conceptual structure seem especially helpful for producing useful syntheses.

The *assumptions made about the audience* are basic to preparing a synthesis document and will influence all other characteristics of the conceptual structure. To a great extent they determine the *intended function* of the synthesis. The *adequacy* of the conceptual structure is closely related to the previous two variables. Are the factors and relationships potential users have to consider in making decisions and taking action all taken into account? Both presenting too much knowledge, organized in an irrelevant way, and too little knowledge, ignoring important variables, can be detrimental to achieving an adequate synthesis. Similarly, the *complexity* of the knowledge can either be so high as to make it difficult to use or simplify situations to such an extent as to be of little value. The last two variables describing the conceptual structure are the *degree of integration* and the *amount of innovation or creativity* characterizing the synthesis.

However, as WARD and REED (1983) indicate, whether a synthesis will be used at all might depend very much on the actual physical structure of the document: its general format, language and readability should match the expectations of potential consumers who might otherwise not accept the information. According to MINTZ (1983) little is known about how evidence affects judgements and behaviour. Therefore these questions should be examined more scrupulously, a view shared by KLEIN (1989), who feels that to improve knowledge syntheses one will have to identify what the most effective synthesis processes and products are.

STRIKE and POSNER (1983) present two models on how research has an effect on practice. The *pipeline model* conceptualizes the dissemination of useful knowledge as a

process of direct contact between the producer and consumer, i.e. the knowledge is targeted to a known audience, need or question. In the *diffusion model* the connection between producer and consumer is indirect, the knowledge is developed without a specific user or application in mind, affecting research, practice and training of practitioners through gradual changes in climate, beliefs and opinions.

Obviously, focusing more specifically on the user's knowledge, skills, needs and settings will increase the practical relevance of a synthesis document. However, this may imply that a given synthesis cannot be useful to a broad audience, the specificity of needs and situations restricting its relevance. In formulating a report writers usually have some idea of who is intended to read their article. It would seem selfevident that they structure and formulate the report accordingly. Meta-analysts do not appear to attach great importance to this issue. Rather, they seem to expect their readers to acquire the knowledge necessary to be intelligent consumers and critics of meta-analyses as BANGERT-DROWNS (1986) explicitly demands (cf. introduction). In effect, this means that the task of achieving effective communication is given to the receiver, rather than regarding it as the sender's problem, which seems contrary to what communication is all about.

Perhaps meta-analysts need to clarify more specifically the functions their approach should fulfil, although after reading various methodological articles (cf. section 2.1) one would imagine that these are clear enough. Yet, while GLASS (1983) characterizes meta-analysis as seeking general conclusions and good generalizations, remaining evaluative and atheoretical in the process, assessing rather than explaining, HEDGES (1986) analyzes why meta-analyses fail as explanations. He postulates that research reviews must succeed as 'explanations' if they are to be useful. In his view explanation consists of more than generalization and summary. It has to make linkages to background beliefs, theories and empirical data clear, and relate the phenomenon to other ideas presumably understood and perceived as relevant by the recipients of the explanation. Thus, apparently HEDGES (1986) considers that explanation is a function meta-analytic reports should fulfil and that the potential consumer determines whether this has succeeded. He focuses on researchers as intended recipients, suggesting that the constructs of treatment, control and outcome should be fairly narrow and specific to the domain being integrated, if meta-analyses are to be more credible. In addition, the quality of the studies being integrated should be discussed with particular attention to the specific difficulties experienced in the domain. Furthermore, qualitative information not explicitly coded as between study difference should not be overlooked as it could play an important role in interpreting results, concluding that perhaps meta-analyses should be made to look more like conventional narrative reviews.

These reflections evidently make use of many of the ideas discussed in connection with knowledge synthesis in general and the advantages of combining qualitative and

quantitative reviewing approaches to improve the usefulness of meta-analytic reviews. Similar considerations could be developed with practitioners as the prospective recipients, making the report more suitable to their presumed objectives. Thus, perhaps, just quantitatively summarizing and integrating research evidence is not enough. It neither reduces the confusion and heterogeneity present in research efforts nor points to a way out of the dilemma.

Whereas the meta-analytic approach was developed to confront the inconsistencies in the findings of primary research and narrative reviews, the quantitative research syntheses themselves are now reaching different conclusions (BRYANT & WORTMAN, 1984; WANOUS, SULLIVAN & MALINAK, 1989). However, as compared to narrative reviews these differences can be identified more readily and completely (ABRAMI et al., 1988). These inconsistencies as well as the hope of achieving a better understanding of the factors influencing research outcomes (CORDRAY & ORWIN, 1983; KULIK, 1984) have led to the trend of publishing reviews of reviews which will be described in the following section.

7.4 Reviews of reviews -- meta-synthesis?

Arguments about the meaningfulness and integrity of quantitative syntheses can potentially be settled through routine critiques and reanalyses (CORDRAY & ORWIN, 1983). However, replication attempts are likely to lead to disagreements. These arise because different sets of studies were used, different comparisons within primary studies were considered worth analyzing, different analysis strategies were applied or investigators examined the data from a different theoretical perspective (ABRAMI et al., 1988; BRYANT & WORTMAN, 1984; DRINKMANN, 1990; WANOUS, SULLIVAN & MALINAK, 1989). Apart from replications of meta-analyses, independent syntheses on similar topics are also available because authors were either not aware that these had been conducted or preferred not to mention or refer to them for some reason.

Researchers hope to resolve inconsistent review findings by integrating them. Introducing the term meta-synthesis for this effort, BRYANT and WORTMAN (1984) promote the cumulation and comparison of conclusions drawn from independent meta-analyses using different criteria to select and evaluate evidence covering the same research domain, feeling that this is the only way in which one can ultimately converge on the truth. On the other hand, researchers support conducting reviews of reviews across different content areas in the hope of finding general rules about factors influencing research (CORDRAY & ORWIN, 1983; KULIK, 1984; PILLEMER & LIGHT, 1980). Another reason given for reviewing reviews was that the literature is so extensive on certain topics that it did not seem feasible to synthesize the primary studies (HAERTEL &

WALBERG, 1980). These intentions are very similar to the purposes of the meta-analytic approach. The main difference appears to be that, instead of using primary studies as database, reviews or meta-analyses now serve this purpose. Quite a number of such documents seem to have been published to date. The impressions reported here are neither unbiased nor necessarily representative as they are based on a small sample that was obtained by chance during the search for literature described in section 6.1.1.

One would imagine that meta-analysts, after stressing the need to improve the methodology of reviewing practices, would translate or extrapolate their systematic approach to this reviewing enterprise. The remarkable fact is, however, in undertaking these efforts meta-analysts seem to regress to the pre-meta-analytic era, forgetting all they have propagated about a rigorous and systematic approach to scientific reviewing. These reviews not only reduce the amount of substantive information given even more than meta-analyses, but also tend to neglect one or more of such elementary matters as representativeness of sampling, exhaustiveness of the literature search or critical evaluation of the meta-analyses being integrated. The quality of documentation leaves much to be desired. If replicability were to be an indication of quality in these cases, an evaluation would frequently have to lead to the pronouncement of a fairly devastating sentence. Whether these reviews of reviews could be useful, is difficult to determine. If, as WALBERG (1984b) suggests, one has to read the original literature to understand the effects of specific factors and conditions, one wonders whether it would not be more effective to just publish the bibliography, thereby considerably reducing the amount one would have to read before having to consult the original sources anyway.

The group of researchers working in concert with Walberg has apparently been publishing reviews of reviews since 1979. These efforts were undertaken largely to find support for the theoretical models of educational productivity developed by Walberg and Hattie. It came as a surprise, therefore, that KULIK and KULIK (1989) do not mention or refer to these syntheses of reviews and meta-analyses in their work, especially as the ground covered is very similar. Furthermore, regarding the publications of the Walberg group, from 1984 onward the documents seem to repeat in varying degrees of detail the work, tables and conclusions already presented previously, one of the more comprehensive summaries being the article by WALBERG (1986), the latest being published by FRASER (1989). In the course of this process, it becomes increasingly difficult to comprehend exactly what sample of studies, reviews and meta-analyses the syntheses are based on.

What sort of general conclusions do these reviews reach, apart from the more specific educational aspects examined? WALBERG, SCHILLER and HAERTEL (1979) conclude that certain conditions and methods consistently produce certain outcomes, but no single method or set of conditions is superior on all outcomes. WAXMAN and WALBERG (1982) find that the reviewers' conclusions are statistically consistent with

one another. ANDERSON (1983) states that the meta-analytic technique has a great deal of stability, being robust to variations in definitions, procedures, sampling and coding. BORGER, LOH, OH and WALBERG (1985) report that the conclusions of primary studies and reviews they integrated reinforce one another. FRASER, WALBERG, WELCH and HATTIE (1987) conclude that the findings across meta-analyses of the same topic were similar, though the pools of sampled studies differed considerably. In the same vein FRASER (1989) indicates that the findings of meta-analyses were predominantly modest in size but surprisingly generalizable, studies yielding similar results in national and international samples, using different research methods, educational subject matter and student populations. KULIK and KULIK (1989) formulate two broad lessons taught by their review of meta-analyses: most experimental treatments and innovative programs in education yield small to moderate positive effects and the mediating influence of design or publication features on these results appear to be quite small, although they can be moderate in some research areas.

Whether consumers find these documents useful will probably depend on the kind of information they are looking for. One thing is certain, however, these articles cannot serve as the sort of normative or actuarial database envisaged by CORDRAY and ORWIN (1983). Nor can these attempts replace or make previous publications superfluous, a hallmark suggested as characteristic of important scientific work (cf. section 2.2). The amount of informed detail is too limited to present a comprehensive picture of the issues crucial to the field being integrated. Perhaps part of the problem is that journal editors allow too little space for reporting results in adequate detail as WALBERG (1986) indicates. On the other hand, perhaps the aims reviewers have set themselves are unrealistic, consequently the standards by which their work is measured are also too high. The nature of behavioral sciences makes it difficult to draw definite conclusions (ANDERSON, 1983), so perhaps one should not expect wonders from the meta-analytic approach to reviewing. Whereas meta-analytic reports definitely cannot solve this problem of the social sciences, they could furnish readers with an organized and comprehensive overview of the theory and findings in particular research domains, using the quantitative analyses to indicate specific trends. It seems unwise to create the impression that concrete conclusions could be achieved, for this can only lead to disenchanting consumers with an approach that has decided advantages over the traditional way of reviewing, if the critiques and recommendations so frequently voiced were heeded to a greater extent.

7.5 Concluding remarks

One of the major problems confronting both practitioners and researchers is the absolute information-overload they have to cope with. For instance, with regard to meta-analytic techniques FRICKE and TREINIES (1985) state that it will be almost impossible

even for experts to register and assimilate all the methodological proposals being published or suggestions for their application. For practitioners the problem of keeping up with scientific developments is even greater. For this reason efforts to accumulate, consolidate and disseminate knowledge are extremely important and can only be encouraged. However, two things are apparent with regard to these efforts. Firstly, reviews or meta-analyses are usually not comprehensive enough to provide an adequate basis for scientists or practitioners to work from. Secondly, they are not stemming the flood of articles one has to cope with, as reading the original or additional articles is still necessary to obtain an accurate impression of the theoretical domain being integrated. Furthermore, in the course of doing so, readers will find that quite a number of these do not report decidedly new information and are repetitive in character: slightly modified versions of articles by the same author published in different journals or books, yet frequently making no references to the similarity, or research reports covering similar ground, yet making no cross-references to other scientists also working in the area.

This state of affairs is neither new nor restricted to the field of meta-analysis (e.g. cf. BRACEY, 1989; KLICHE, 1990; MEINEFELD, 1985; STRAUSS, 1969; STRINGFIELD, 1991). A number of factors contribute to it, some of which appear to be difficult to alter. One of these is the apparent pressure on scientists to write and publish articles, come what may. This is the main measure of their scientific worth, but it can also result in sloppy publishing practices. Closely related to the previous aspect is the competitive nature of academic work and the existence of citation 'syndicates'. Both are not conducive to co-ordinating research efforts, placing the work within a larger perspective and making interconnections clear through explicit and comprehensive cross-references to previous theory, research and ideas. Furthermore, apparently primary research and theory development are considered more prestigious scientific enterprises, allegedly requiring higher and more refined intellectual capacities than efforts to consolidate and synthesize existing knowledge (theory and research). Yet, as one can deduce from the evaluation and discussion given above, this is a complex undertaking, high quality being difficult to achieve.

Although such efforts should be a prerequisite and the basis of all further scientific developments, they are apparently often regarded as mere drudgery and of secondary importance, thus handled in a cursory way. This constitutes a factor that can and should be changed. Information technology has simplified the task of finding literature relevant to specific topics, but the databases are usually incomplete, the descriptors for searching them are often fairly general, limiting the precision of the search and the terminological diversity prevalent in the social sciences does not help matters. Although one can hardly hope to gain access to most or all sources of information relevant to a topic one should at least try to incorporate most of what is available. Particularly when summarizing the state-of-the-art or integrating research evidence special efforts need to be made to provide a comprehensive overview of the theory and empirical results, perhaps even provid-

ing readers with references to related topics, research or articles not specifically discussed. These practices could possibly reduce the amount of additional reading necessary for those not well versed in the particular domain, simplify the task of identifying literature relevant to related questions and provide other scientists with a good foundation to build on in continuing research efforts. The value of good bibliographies and detailed narrative or meta-analytic reviews should not be underestimated by either practitioners or scientists.

As indicated above, the potential value of meta-analyses could be increased, if these matters were given more attention, making them more informative as well as useful. Largely, the question of usefulness tends to be neglected. Focusing on it more specifically would enhance the practical relevance of meta-analyses for both scientists and practitioners. The degree of practical relevance scientific endeavours can hope to attain is necessarily limited by the nature of social science and practice. But up to now even these possibilities have not been fully exploited. According to GOOD (1983a) researchers should think more seriously about the meaning and relation of findings to practice, a sentiment that can also be applied to meta-analytic research. How matters could be improved in general, regardless of whether fellow-scientists or practitioners constitute the intended audience, has been briefly outlined above.

More specifically for the benefit of practitioners, it might also be important to clarify what meta-analyses, research or theory can realistically achieve. This could help to avoid disappointing expectations or prevent consumers from overreacting to research findings and considering results too narrowly as GOOD (1983b) fears. Ideas concerning this issue are repeatedly voiced by GOOD (1979, 1983a, 1983b) in his reviews on educational research and appear worth noting. He points out that theory and research cannot solve problems or provide answers. They can, however, clarify issues by providing guidelines, concepts, frames of reference or directions to think about situations; specify dimensions relevant to understanding phenomena; yield an awareness of problems and alternative ways of solving or responding to them by extending the range of hypotheses and alternative strategies considered; or sensitize practitioners to possible consequences of actions.

Furthermore, Good stresses that it is difficult to translate findings into recommendations or a set of specific behavioral prescriptions. On the one hand the knowledge is usually too limited. Whereas the variables affecting teaching and learning are numerous, complex and interrelated, research focuses mainly on single variables rather than more comprehensive contexts. On the other hand the nature of problems and situations to which practitioners want to apply findings are varied and unique. The generalizations derived from research and theory are, however, indeterminate in the sense that they cannot predict what will happen in a particular case. Consequently, blind applications of research findings should be discouraged. Practitioners should rather be encour-

aged to adapt findings to their own situation and monitor their impact. Therefore GOOD (1983a, 1983b) propagates presenting knowledge or information in a decision making format along with judgemental skills that help persons to examine concepts and apply them to their unique settings, concluding that the best research can hope to achieve is to help practitioners analyze their own settings and become more adept at seeing, understanding and reacting to conditions they face.

These reflections seem to sum up what meta-analysts can realistically hope to achieve with regard to practical relevance for practitioners. Most of these ideas link up to the suggestions made for improving the quality of meta-analyses. They also particularly underline the fact that little is to be gained in the way of understanding concepts or how research findings can be translated into practice by presenting predominantly quantitative data, or effect sizes accurate to a few decimal points, without also providing sufficient information on the theoretical background, allowing readers to recognize the theoretical and practical implications and intricate interrelationships of the findings.

In summary one may conclude that the advent of meta-analysis and the resulting discussion of reviewing practices in general have led to some improvements. As yet, however, one should not be satisfied with what has been achieved, but rather intensify efforts to improve the methods of knowledge synthesis and accumulation both by attending more specifically to those issues identified as enhancing their quality and by continuing research on what could possibly improve their quality and usefulness.

GLOSSARY

(italics indicate that the word is described elsewhere in the glossary; page references to the theoretical sections of the text are given for terms of special importance)

ANOVA: short for ANalysis Of VAriance; refers to statistical procedures based on the separation of a sample's total *variance* into components associated with defined sources of variation; the aim is to find out whether there are differences between various components so as to identify sources of variations

alternative hypothesis: commonly used to denote the statistical hypothesis formulated in contrast to the *null hypothesis*, postulating that the variables under investigation are related or that real differences between experimental groups exist; also used in the sense of *rival hypothesis*

assignment rules: rules by which subjects are assigned to various experimental groups, e.g. random assignment (chance), self-assignment, matching

binomial effect size display: (BESD) a method developed to make the practical import of *effect sizes* evident and their interpretation more transparent; they are transformed into a measure of success rate, indicating the improvement or difference that can be expected, e.g. a *correlation* of 0.3 is considered a moderate *effect size*, indicating that 9% of the variance is explained, in terms of the BESD it means that the success rate has changed from 35% to 65%, i.e. has practically been doubled (p.25)

blocking: has the general purpose of dividing material into blocks which are homogeneous; a term deriving from factorial experiments; a *control technique* in which blocks are used to isolate sources of heterogeneity; uncontrolled variation is measured by comparisons between blocks

box count: also referred to as *vote count*; technique in which the number of neutral, positive and negative statistical hypothesis test results in the sample of reviewed studies is determined with the aim of establishing whether one of the categories predominates, indicating whether the research results in general appear to support the research hypothesis or not

capitalization on chance: results from a misuse of the principle of significance testing; when many tests are performed on the same data-set the likelihood of finding apparently *significant results* is increased; one capitalizes on the occurrence of a few extreme cases among many comparisons (p.27, 47)

coding: process in which a study is quantified by recording the information it contains according to a specific system of categories covering all study features and characteristics relevant to the meta-analytic research hypothesis so that the coding represents a comprehensive, quantitative description of the respective study

coding reliability: measure of the degree of agreement achieved when several persons independently code the same article; also see *intercoder agreement/reliability* (p.21, 29, 46)

combined probability: if several independent tests of the same research hypothesis exist, the probability with which their respective *null hypotheses* are false can be combined by a variety of techniques to obtain a sounder test of the research hypothesis (p.18, 25)

common metric: also referred to as common scale; means that values are directly comparable because they are expressed in the same dimension; see *standardizing unit* (p.11, 14, 22, 24)

confounding: occurs in an experiment if a variable is systematically related to the *independent variable* and may differentially affect the *dependent variable*, thus making it impossible to interpret the results; the potential effects of the variables under investigation cannot be separated from the possible effects of other variables; to avoid this there are a variety of *control techniques* that can be employed, depending on the experimental design

construct definitions: construct is another term for theoretical concept; refers to a characteristic that cannot be observed or measured directly (e.g. intelligence); it is defined in terms of other theoretical concepts and its existence can only be established by specifying theoretically deduced indicators that can be measured directly; also see *operationalization* (p.14, 22, 28)

construct validity: concerns the adequacy with which theoretical concepts have been translated into observable variables; the extent to which operations reflect the research constructs; the approximate validity of generalizing about the *constructs* from the research operations; thus, **construct validity of causes** concerns the question of whether the experimental manipulations adequately represent the particular theoretical cause concepts (*independent variables*); **construct validity of effects** refers to whether the procedures used to measure the effects (*dependent variable*) actually tap the theoretical factors they are meant to (p.31, 34, 39, 41)

control techniques: are procedures employed in experiments to help ensure that the results are valid, i.e. can be interpreted unambiguously; they are an important aspect in designing experiments; examples of control techniques are elimination, constancy of conditions, balancing and randomization

correlation: a measure of relationship between two or more variables; variables are correlated if their values covary, i.e. values of the one are systematically related to either equally high values (positive correlation) or correspondingly low values (negative correlation) of another

dependent variable: is the variable presumed to be affected by manipulations of the *independent variable*; changes in dependent variables are measured to determine the effect of independent variables; the value of dependent variables can be predicted from values of the independent variable

effect size: expresses the degree to which a phenomenon or relationship is present or manifested in a population; a measure of the magnitude of an experimental effect; many different formula are available to calculate it; the most commonly used are: the difference between two group means divided by their common standard deviation (d) and the correlation (r) between two continuous variables (p.11, 14, 24-25, 46); **interpretable effect size:** are calculated using uncorrected measures of the *dependent variable* (cf. *final status score*) and the *standard deviation* as *standardizing unit*; because of this they are conceptually equivalent and can be interpreted on a *common scale* which is a prerequisite for making valid conclusions about effects in meta-analytic studies (p.24); **operative effect sizes:** are calculated using adjusted values, removing sources of irrelevant variation; they are not conceptually equivalent when calculated for different experimental designs because different *standardizing units* are used and thus cannot be interpreted in a single way; their use in meta-analyses is not recommended because they are not necessarily directly comparable (p.24); **interpretation of effect sizes:** the magnitudes of effect sizes are most frequently interpreted as standard deviation difference between two groups or percentiles (the percentage of persons in the one group exceeding the scores of those in the other group); Cohen's classification of large (d=.8, r=.5), moderate (d=.5; r=.3) and small (d=.2; r=.1) effect sizes for the area of social sciences is a relative interpretation; what is labeled small or large should depend on the magnitude relative to a variety of related estimates chosen because of their substantive relevance to the topic under study; behavioral equivalents such as gains on standardized tests are also used to make the meaning of effect sizes clear (p.25, 44); also see *binomial effect size display*

external validity: concerns the question of whether the results of the investigation are applicable to situations outside the immediate confines of the study; the extent to which results can be generalized to and across alternative measures of cause and effect or different types of persons, settings and times; to a large degree it concerns the adequacy of the sample, its *representativeness* of persons, settings, times and constructs (p.13, 31, 34-35, 39, 41)

fail-safe-test: estimates the number of studies with nonsignificant results that would have to be added to the retrieved ones in order to change the conclusion that *significant effects* exist, i.e. the *combined probability* of the results is no longer significant (p.18)

file drawer study / problem: a file drawer study is one that has not been published; these are not easily retrieved to be included in meta-analyses; this becomes a problem if the reason for their not being published is that their results were not significant; journals might be publishing an overproportional number of studies with significant results; if they are predominantly integrated by meta-analyses it could falsify the general picture; a test of the possible extent of this problem is provided by the *fail-safe test* (p.18, 47)

final status score: is the actual measure obtained for the *dependent variable* without any corrections; in contrast derived measures such as gain scores are corrected or adjusted for pretreatment differences; compare interpretable and operative *effect size* (p.25)

heterogeneous samples: subjects of the sample differ with respect to various characteristics; sample is composed of diverse constituents

homogeneous samples: are composed of similar constituents; subjects of the sample share many characteristics, e.g. age, sex, race, social class

independent variable: also referred to as explanatory or experimental variable, i.e. the variable used to predict the level of a *dependent variable* or manipulated in an experiment to cause an effect in some dependent variable

inflated N: artificially increased sample size, e.g. by including non-independent observations or measures as elements of the sample; negatively affects the *validity* of the conclusions that can be drawn from statistical tests (p.24)

instrumentation problems: arise if the sensitivity or accuracy of the measurement instrument changes during the course of the study; the *reliability* of measurement is essential to guarantee the *validity* of results

interaction effect: implies the presence of some mediator/ moderator; indicates the extent to which the effect of a variable depends on the presence or absence of some other variable; term derives from factorial experiments; effect is attributable to a combination of variables; compare *main effect*

intercoder agreement / reliability: degree of agreement in the recordings/ratings/calculations/coding of several independent coders/raters; see *coding reliability*

internal validity: deals with the relations between research operations; the extent to which the existence of causal relationships can be deduced or the absence of causal relations implies the absence of cause; concerns the question of whether the particular treatment or manipulations produced the effects or whether plausible *rival hypotheses* exist to explain these; largely depends on the adequacy of experimental designs (p.31, 32, 39, 45)

intervening variables: term used to account for internal, directly unobservable psychological processes that account for some form of behaviour

main effect: the effect of an experimental variable measured while ignoring the effect of other variables that also form part of the experiment by holding them constant across all levels of the main variable; term derives from factorial experiments; compare *interaction effect*

mean: commonly referred to as average; is equal to the arithmetic sum of all the individual values of the observations divided by the number of the observations

median: midmost measure of a set of observations; the number of observations with values below and above this measure is equal

mediating variables: see *moderating variables*

missing data: in primary studies is a problem that can threaten the *validity* and representativeness of a meta-analysis because the information needed to code relevant study features or the statistical data necessary to calculate *effect sizes* is not available; this reduces

the sample size and could introduce *systematic bias* if the missing data are related to important characteristics of the primary studies (p.18, 19, 46; also see microlevel reporting, p.30)

moderating variables: influence the strength of the relationship between two or more other variables; also referred to as *mediating variables*

non-independence: when the value of an observation or the occurrence of an event is determined or influenced by the value or presence of some other event; most statistical tests assume that events are independent, if they are not the analysis is invalidated and difficult to interpret; also see *inflated N* (p.24, 47)

null hypothesis: the particular statistical hypothesis under test; usually postulates that the variables under investigation are unrelated or that there are no differences between experimental groups except due to chance; relates to *Type I and Type II errors*

operationalization: translation of theoretical variables into empirical terms by specifying procedures and instruments, i.e. operations to be performed in order to be able to manipulate or measure a *construct*

outliers: observations in a sample that are so different from the remainder that they might derive from some other population or are the result of some fault

parameter: is a theoretical, population value estimated on the basis of a sample *statistic* for the total population or universe from which the sample was drawn

power of a statistical test: is the probability that it rejects the *null hypothesis* when the *alternative hypothesis* is true, i.e. its sensitivity to detect effects correctly; the power is greatest when the probability of a *Type II error* is least; different statistical tests have different power functions, exhibited as graphs these give a clear picture of the 'performance' of a test

random sample: selected without presence of *systematic bias*; every member of a population has the same specified probability of being included in the sample, usually means equal chance

regression analysis: statistical method to investigate the relationships between variables

reliability: degree to which repeated measurement with an instrument will give the same or similar readings; usually expressed as the degree of *correlation* between the measures (p.12, 29-31, 46, 48)

representative sample: means that it is typical of the population from which it was selected, often implying that all members of the population in question had an equal chance to represent it; compare *random sample*

rival hypothesis: a plausible alternative to the proposed *alternative hypothesis* as explanation of the findings; can arise because *operationalizations* only provide a partial defini-

tion of a theoretical concept or due to inadequate experimental designs; see *control techniques*

sampling plan: all steps taken in selecting a sample from a given population; set of rules for drawing a sample in an unequivocal manner

significant results / effects: a result or effect that according to a statistical test is unlikely to have occurred by chance with a predefined probability (level of significance) of making a wrong interpretation; usually the probability is set at .05 or .01 which means that there is a chance of 5 or 1 to a 100 that the effect or result obtained is not real but only due to chance; a test of significance is one which purports to provide a test of the hypothesis that an effect is absent; an effect is called statistically significant if the value of the *statistic* used to test it lies outside acceptable limits; also see *Type I* and *Type II error*

spurious effects/findings: effect not present in the original material but induced by the method of handling; a *correlation* between variables exists even though the original values of the observations are unrelated

standard deviation: the most widely used measure of the spread of sample findings; statistically equal to the square root of the *variance*

standard error: estimates the degree to which the calculated sample *statistic* deviates from the 'true' population *parameter*; statistically the *standard deviation* of a sampling distribution

standardized test: one which has been administered to a *representative sample* of people to establish norms for evaluating the scores and has been subjected to tests of *validity* and *reliability* to ensure its quality

standardizing unit: used to transform observations measured in different dimensions to a *common metric* or scale; a sample's *standard deviation* is often used for this purpose, as it allows observations that were measured in completely different units to be compared as quasi dimensionless quantities; see *effect size*

statistic: a measure calculated from the sample of observations under investigation to summarize some aspect, e.g. its central tendencies (*mean*, *median*) or dispersion (*standard deviation*, *variance*)

statistical conclusion validity: refers to the adequacy of the employed statistical procedures and whether the conclusions based on statistical evidence are correct; whether the methods were correctly applied and sensitive enough to detect effects, should they exist (p.31, 33, 39, 45)

statistical power function: see *power of a statistical test*

stem-and-leaf table: the representation of a large number of *correlations* or *effect sizes* in the form of a table by listing the first significant number followed by the subsequent number values of all observations with the same first significant number, e.g. 2.1, 2.5, 2.0,

2.3, 2.9 would be listed as 2 01359, giving a graphical as well as numerical indication of the distribution of the sample of values

stratifying variable: a characteristic used to divide a sample into subgroups, e.g. sex, age, geographic region etc.

systematic bias: not random, but uniformly related to some other relevant aspect; it consistently and artificially inflates or deflates scores thus invalidating the findings; it can be counteracted through applying *control techniques* and adequate experimental designs

t-test: statistical test based on the t- or Student's distribution; usually used to test whether the *means* of two independent or dependent samples differ more than can be expected by chance

Type I error: occurs when a statistical *null hypothesis* is rejected when it ought to be accepted, i.e. is true; the frequency with which this occurs can be controlled by the appropriate selection of the level of significance; see *significant results*

Type II error: occurs when the *null hypothesis* is not rejected though it is false; it cannot be controlled as easily as the *Type I error*, usually the magnitude of a Type I error is fixed and the Type II error minimized within this restriction; also see *power of a statistical test*

unbiased estimates: term used for a *statistic* estimating a *parameter* value; unbiasedness is a property of the sampling distribution and not strictly a property of a single *statistic*; implies that in the long run the *mean* of such a *statistic* computed from a large number of samples of equal size will be equal to the *parameter*

unit of analysis: refers to the entity that goes into the statistical evaluation, e.g. in educational research the pupil or student, the class or the school can serve as unit of analysis; in meta-analyses it can be the individual findings of the primary studies or the primary studies themselves (p.14, 24, 26, 27, 33, 46, 47)

validity: in general refers to an aspect of the quality of an empirical study, the extent to which the 'truth' of the propositions tested is approximated; concerns the inquiry into the nature and meaning of the variables studied, the adequacy of the experimental design, implementation, statistical analyses and conclusions; there is no one validity, different kinds have to be examined; see *construct*, *external*, *internal* and *statistical conclusion validity* (p.12, 29, 31, 37; also see prior validity and macrolevel reporting, p.35)

variance: a measure of variability; indicates the spread or range of sample findings; statistically the sum of the squares of the deviations from the arithmetic *mean* divided by the sample size; also see *standard deviation*

vote count: see *box count*

weighting: attaching a numerical coefficient to observations so as to reflect their relative importance

References

ABRAMI, P.C., COHEN, P.A. & d'APOLLONIA, S. (1988). Implementation problems in meta-analysis. *Review of Educational Research, 58* (2), 151-179.

ABRAMI, P.C., d'APOLLONIA, S. & COHEN, P.A. (1990). Validity of student ratings of instruction: What we know and what we do not. *Journal of Educational Psychology, 82* (2), 219-231.

ANDERSON, R.D. (1983). A consolidation and appraisal of science meta-analysis. *Journal of Research in Science and Teaching, 20*, 497-509.

BANGERT-DROWNS, R.L. (1986). Review of developments in meta-analytic method. *Psychological Bulletin, 99* (3), 388-399.

BANGERT-DROWNS, R.L., KULIK, J A. & KULIK, C-L.C. (1983). Effects of coaching programs on achievement test performance. *Review of Educational Research, 53* (4), 571-585.

BORGER, J.B., LO, C-l., OH, S-s. & WALBERG, H.J. (1985). Effective schools: A quantitative synthesis of constructs. *Journal of Classroom Interactions, 20* (2), 12-17.

BRACEY, G. (1989). Why so much education research is irrelevant, imitative and ignored. *American School Board Journal, 176* (7), 20-22, 42.

BRINBERG, D. & McGARTH, J.E. (1982). A network of validity concepts within the research process. In D. Brinberg & L.H. Kidder (Eds.), *Forms of validity in research* (pp.5-21). San Francisco: Jossey-Bass.

BRYANT, F.B. & WORTMAN, P.M. (1984). Methodological issues in the meta-analysis of quasi-experiments. In W.H. Yeaton & P.M. Wortman (Eds.), *Issues in Data Synthesis*. New Directions for Program Evaluation (no.24, pp.5-24). San Francisco: Jossey-Bassey.

BULLOCK, R.J. & SVYANTEK, D.J. (1985). Analyzing meta-analysis: Potential problems, an unsuccessful replication, and evaluation criteria. *Journal of Applied Psychology, 70*, 108-115.

CAMPBELL, D.T. (1987). Guidelines for monitoring the scientific competence of preventive intervention research centers. *Knowledge: Creation, Diffusion, Utilization, 8*, 389-430.

CHELIMSKY, E. & MORRA, L.G. (1984). Evaluation synthesis for the legislative user. In W.H. Yeaton & P.M. Wortman (Eds.), *Issues in data synthesis*. New Directions for program evaluation (no.24, pp. 75-89). San Francisco: Jossey-Bassey.

COOK, T.D. & CAMPBELL, D.T. (1979). *Quasi-experimentation, design and analysis for field settings*. Chicago: Rand McNally College Publishing Company.

COOK, T.D. & LEVITON, L.C. (1980). Reviewing the literature: A comparison of traditional methods with meta-analysis. *Journal of Personality*, *48*, 449-472.

COOPER, H.M. (1981). On the significance of effects and the effects of significance. *Journal of Personality and Social Psychology*, *41* (5), 1013-1018.

COOPER, H.M. (1982). Scientific guidelines for conducting integrative research reviews. *Review of Educational Research*, *52* (2), 291-302.

COOPER, H.M. (1984). *The integrative research review: A systematic approach*. Applied social research methods series, Volume 2. Beverly Hills: Sage Publications.

COOPER, H.M. & ARKIN, R.M. (1981). On quantitative reviewing. *Journal of Personality*, *49*, 225-230.

COOPER, H. & RIBBLE, R.G. (1989). Influences on the outcome of literature searches for integrative research reviews. *Knowledge: Creation, Diffusion, Utilization*, *10* (3), 179-201.

COOPER, H.M. & ROSENTHAL, R. (1980). Statistical versus traditional procedures for summarizing research results. *Psychological Bulletin*, *87*, 442-449.

CORDRAY, D.S. & ORWIN, R.C. (1983). Improving the quality of evidence: Interconnections among primary evaluation, secondary analysis and quantitative synthesis. In R.J. Light (Ed.), *Evaluation Studies Review Annual*, (Vol.8, pp.91-119). Beverly Hills: Sage Publications.

DRINKMANN, A. (1990). *Methodenkritische Untersuchungen zur Metaanalyse*. Weinheim: Deutscher Studienverlag.

FISKE, D.W. (1983). The meta-analytic revolution in outcome research. *Journal of Consulting and Clinical Psychology*, *51*, 65-70.

FRASER, B.J. (1989). Research synthesis on school and instructional effectiveness. *International Journal of Educational Research*, *13*, 707-719.

FRASER, B.J., WALBERG, H.J., WELCH, W.W. & HATTIE, J.A. (1987). Synthesis of educational productivity research. *International Journal of Educational Research*, *11*, 145-252.

FRICKE, R. & TREINIES, G. (1985). *Einführung in die Metaanalyse*. Bern: Verlag Hans Huber.

FRICKE, R. (1985). Klarheit der Lehrersprache und Schülerleistung: Eine Metaanalyse. In K. Aurin & B. Schwarz (Hrsg.), *Die Erforschung pädagogischer Wirkungsfelder* (pp.205-211). Freiburg: Arbeitsgruppe für empirische pädagogische Forschung in der DGfE.

FROMM, M. (1990). Zur Verbindung quantitativer und qualitativer Methoden. *Pädagogische Rundschau*, *44*, 469-481.

GLASS, G.V. (1977). Integrating findings: The meta-analysis of research. In L.S. Shulman (Ed.), *Review of Research in Education* (No.5, pp.351-379). Itasca: F.E. Peackock Publishers.

GLASS, G.V. (1983). Synthesizing empirical research: Meta-analysis. In S.A. Ward & L.J. Reed (Eds.), *Knowledge structure and use: Implications for synthesis and interpretation* (pp.399-421). Philadelphia: Temple University Press.

GLASS, G.V. & KLIEGL, R.M. (1983). An apology for research integration in the study of psychotherapy. *Journal of Consulting and Clinical Psychology*, *51*(1), 28-41.

GLASS, G.V., McGAW, B. & SMITH, M.L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

GLASS, G.V. & SMITH, M.L. (1979). Meta-analysis of research on class size and achievement. *Educational Evaluation and Policy Analysis*, *1* (1), 2-16.

GOOD, T. (1979). Teacher effectiveness in the elementary school: What do we know about it now? *Journal of Teacher Education*, *30*, 52-64.

GOOD, T.L. (1983a). Classroom Research: A decade of progress. *Educational Psychologist*, *18* (3), 127-144.

GOOD, T. (1983b). Research on classroom teaching. In L.S. Shulman & G. Sykes (Eds.), *Handbook of Teaching and Policy* (pp.42-80). New York: Longman.

GREEN, B.F. & HALL, J.A. (1984). Quantitative methods for literature reviews. *Annual Review of Psychology*, *35*, 37-53.

GUSKIN, S.L. (1984). Problems and promises of meta-analysis in special education. *Journal of Special Education*, *18* (1), 73-80.

HAERTEL, E.H. & WALBERG, H.J. (1980). Investigating an educational productivity model. In H.J. Walberg & E.H. Haertel (Eds.), *Research integration: The state of the art*. Evaluation in Education: An International Review Series, 4 (1), 103-104. New York: Pergamon Press.

HAGER, W. (1984). Metaanalyse: Zahlen als Psychologieersatz? *Psychologie, Erziehung und Unterricht*, *31*, 64-70.

HAGER, W. (1985). Beurteilung inhaltlicher Hypothesen als Alternative zur Metaanalyse: Wirken Zielvorgaben aufmerksamkeitslenkend oder allgemein motivierend? *Psychologische Beiträge*, *27*, 200-217.

HANSFORD, B.C. & HATTIE, J.A. (1982). The relationship between self and achievement / performance measures. *Journal of Educational Research*, *52* (1), 123-142.

HEDGES, L.V. (1980). Unbiased estimation effect size. *Evaluation of Educational Research*, *4*, 25-27.

HEDGES, L.V. (1982a). Estimation of effect size from a series of independent experiments. *Psychological Bulletin*, *92* (2), 490-499.

HEDGES, L.V. (1982b). Fitting categorical models to effect sizes from a series of experiments. *Journal of Educational Statistics*, *7*(2), 119-137.

HEDGES, L.V. (1983). A random effects model for effect sizes. *Psychological Bulletin*, *93* (2), 388-395.

HEDGES, L.V. (1984). Advances in statistical methods for meta-analysis. In W.H. Yeaton & P.M. Wortman (Eds.), *Issues in data-synthesis*. New directions for program evaluation (no.24, pp.25-42). San Francisco: Jossey-Bassey.

HEDGES, L.V. (1986). Issues in meta-analysis. In E.Z. Rothkopf (Ed.), *Review of research in education* (no.13, pp.353-398). Washington, DC: American Educational Research Association.

HEDGES, L.V. & OLKIN, I. (1985). *Statistical methods for meta-analysis*. Orlando: Academic Press Inc.

HORNKE, L. (1983). Integration empirischer Forschungsergebnisse? Zum Problem der vorstrukturierenden Lernhilfen im Sinne Ausubels. *Psychologie in Erziehung und Unterricht*, *30*, 54-63.

HUNTER, J.E., SCHMIDT, F.L. & JACKSON, G.B. (1982). *Meta-analysis: Cumulating research findings across studies*. Studying organizations: Innovations in methodology, Vol.4. Beverly Hills, CA: Sage.

JACKSON, G.B. (1980). Methods for integrative reviews. *Review of Educational Research*, *50* (3), 438-460.

JOHNSON, D.W., MARUYAMA, G., JOHNSON, R., NELSON, D. & SKON, L. (1981). Effects of cooperative, competitive and individualistic goal structures on achievement: A meta-analysis. *Psychological Bulletin*, *89* (1), 47-62.

JUDD, C.M. & KENNY, D.A. (1982). Research design and research validity. In D. Brinberg & L.H. Kidder (Eds.), *Forms of validity in research* (pp.23-39). San Francisco: Jossey-Bass.

KAVALE, K.A. (1988). Using meta-analysis to answer the question: What are the important, manipulable influences in school learning? *School Psychology Review*, *17* (4), 644-650.

KEMERY, E.R., MOSSHOLDER, K.W. & DUNLAP, W.P (1989). Meta-analysis and moderator variables: A cautionary note on transportability. *Journal of Applied Psychology*, *74* (1), 168-170.

KEMERY, E.R., MOSSHOLDER, K.W. & ROTH, L. (1987). The power of the Schmidt and Hunter additive model of validity generalization. *Journal of Applied Psychology*, *72*, 30-37.

KENDALL, P.C. & MARUYAMA, G. (1985). Meta-analysis: On the road to synthesis of knowledge? *Clinical Psychology Review*, *5*, 79-89.

KIESLER, C.A. (1985). Meta-analysis, clinical psychology, and social policy. *Clinical Psychology Review*, *5*, 3-12.

KLAUER, K.J. (1981). Zielorientiertes Lehren und Lernen bei Lehrtexten. *Unterrichtswissenschaft*, *4*, 300-318.

KLAUER, K.J. (1984). Intentional and incidental learning with instructional texts: A meta-analysis for 1970 – 1980. *American Educational Research Journal*, *21* (2), 323-339.

KLEIN, S.S. (1989). Research and practice: Implications for knowledge synthesis in education. *Knowledge: Creation, Diffusion, Utilization*, *11* (1), 58-78.

KLICHE, T. (1990). Das zweite Gesicht der Macht. Bibliographie zur Politischen und politotropen Psychologie des Mythos. *PP-Aktuell*, *9* (3), 122-157.

KRAEMER, H.C. & ANDREWS, G. (1982). A nonparametric technique for meta-analysis effect size calculation. *Psychological Bulletin*, *91* (2), 404-412.

KULIK, J.A. (1984). *The uses and misuses of meta-analysis*. Paper presented at the meeting of the American Educational Research Association, New Orleans. ERIC Document No. ED 248618.

KULIK, J.A. & KULIK, C.L.C. (1986). *Operative and interpretable effect sizes in meta-analysis*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco. ERIC Document No. ED 276759.

KULIK, J.A. & KULIK, C.L.C. (1988). *Meta-analysis: Historical origins and contemporary practice*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans. ERIC Document No. ED 297015.

KULIK, J.A. & KULIK, C.-L.C. (1989). Meta-analysis in education. *International Journal of Educational Research*, *13* (3), entire issue.

LAMNEK, S. (1988). *Qualitative Sozialforschung*. Band 1: Methodologie. München: Psychologie Verlags Union.

LEVITON, L.C. & COOK, T.D. (1981). What differentiates meta-analysis from other forms of review. *Journal of Personality*, *49*, 231-236.

LIGHT, R.J. & PILLEMER, D.B. (1982). Numbers and narrative: Combining their strengths in research reviews. *Harvard Educational Review*, *52* (1), 1-26.

LIGHT, R.J. & PILLEMER, D.B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard Unversity Press.

LINDBLOM, C.E. (1987). Alternatives to validity. *Knowledge: Creation, Diffusion, Utilization*, *8*, 509-520.

MATT, G.E. (1989). Decision rules for selecting effect sizes in meta-analysis: A review and reanalysis of psychotherapy outcome studies. *Psychological Bulletin*, *105* (1), 106-115.

McGAW, B. & GLASS, G.V. (1980). Choice of the metric for effect size in meta-analysis. *American Educational Research Journal*, *7*, 325-337.

MEEHL, P.E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46* (4), 806-834.

MEINEFELD, W. (1985). Die Rezeption empirischer Forschungsergebnisse – eine Frage von Treu und Glauben? *Zeitschrift für Soziologie*, *14* (4), 297-314.

MINTZ, J. (1983). Integrating research evidence. A commentary on meta-analysis. *Journal of Consulting and Clinical Psychology*, *51*, 71-75.

MOORE, D.W. & READENCE, J.E. (1984). A quantitative and qualitative review of graphic advance organizer research. *Journal of Educational Research*, *78* (1), 11-17.

NOBLIT, G.W. & HARE, R.D. (1988). *Meta-ethnography: Synthesizing qualitative studies*. Newbury Park, CA: Sage Publications.

ORWIN, R.G. & CORDRAY, D.S. (1985). Effects of deficient reporting on meta-analysis: A conceptual framework and reanalysis. *Psychological Bulletin*, *97*, 134-147.

PILLEMER, D.B. & LIGHT, R.J. (1980). Synthesizing outcomes: How to use research evidence from many studies. *Harvard Educational Review*, *50* (2), 176-195.

RAUDENBUSH, S.W. (1983). Utilizing controversy as a source of hypotheses for meta-analysis. In R.J. Light (Ed.), *Evaluation Studies Review Annual* (Vol.8, pp.303-325). Beverly Hills: Sage Publications.

RAUDENBUSH, S.W. (1984). Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy induction: A synthesis of findings from 18 experiments. *Journal of Educational Psychology*, *76* (1), 85-97.

RAUDENBUSH, S.W. & BRYK, A.S. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics*, *10* (2), 75-98.

REICHARDT, C.S. & COOK, T.D. (1978). Beyond qualitative versus quantitative methods. In T.D. Cook & C.S. Reichardt (Eds.), *Qualitative and quantitative methods in evaluation research* (pp.7-32). Beverly Hills: Sage Publications.

RESTIVO, S. & LOUGHLIN, J. (1987). Critical sociology of science and scientific validity. *Knowledge: Creation, Diffusion, Utilization*, *8*, 486-508.

ROSENTHAL, R. (1978). Combining results of independent studies. *Psychological Bulletin*, *85* (1), 185-193.

ROSENTHAL, R. (1979). The „File Drawer Problem" and tolerance for null results. *Psychological Bulletin*, *86* (3), 638-641.

ROSENTHAL, R. (1982). Valid interpretations of quantitative research results. In D. Brinberg & L.H. Kidder (Eds.), *New Directions for Methodology of Social and Behavioral Science: Forms of validity in research* (no.12, pp.59-75). San Francisco: Jossey-Bass.

ROSENTHAL, R. (1984). *Meta-analytic procedures for social research*. Applied social research methods series, Volume 6. Beverly Hills: Sage Publications.

ROSENTHAL, R. & RUBIN, D.B. (1982a). Comparing effect sizes of independent studies. *Psychological Bulletin*, *92* (2), 500-504.

ROSENTHAL, R. & RUBIN, D.B. (1982b). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, *74* (2), 166-169.

ROSENTHAL, R. & RUBIN, D.B. (1982c). Further meta-analytic procedures for assessing cognitive gender differences. *Journal of Educational Psychology*, *74*(5), 708-712.

ROSENTHAL, R. & RUBIN, D.B. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin*, *99* (3), 400-406.

SACKETT, P.R., HARRIS, M.M. & ORR, J.M. (1986). On seeking moderator variables in the meta-analysis of correlational data: A Monte Carlo investigation of statistical power and resistance to Type I error. *Journal of Applied Psychology*, *71*, 302-310.

SACKS, H.S., BERRIER, J., REITMAN, D., ANCONA-BERK, V.A. & CHALMERS, T.C. (1987). Meta-analyses of randomized controlled trials. *New England Journal of Medicine*, *316*, 450-455. [cited in DRINKMANN, 1990]

SCHÖNEMANN, P.H. (1990). Burt Rätsel – Kommentar zu Wottawa: Einige Überlegungen zu (Fehl-) Entwicklungen der psychologischen Methodenlehre. *Psychologische Rundschau, 41*, 103-105.

SCRUGGS, T.E., MASTROPIERI, M.A. & CASTO, G. (1987). The quantitative synthesis of single-subject research: Methodology and validation. *Remedial and Special Education, 8* (2), 24-33.

SECHREST, L. & YEATON, W. (1981). Empirical bases for estimating effect size. In R.F. Boruch et al. (Eds.), *Reanalyzing Program Evaluations* (pp.212-224).San Francisco: Jossey-Bass.

SLAVIN, R.E. (1984). Meta-analysis in education: How has it been used? *Educational Researcher, 13* (8), 6-15.

SLAVIN, R.E., (1986). Best evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational Researcher, 15*, 5-11.

SMITH, M.L. & GLASS, G.V. (1980). Meta-analysis of research on class size and its relationship to attitudes and instruction. *American Educational Research Journal, 17* (4), 419-433.

SPECTOR, P.E. & LEVINE, E.L. (1987). Meta-analysis for integrating study outcomes: A Monte Carlo study of its susceptibility to Type I and Type II errors. *Journal of Applied Psychology, 72*, 3-9.

STANLEY, J.C. (1987). Note about possible bias resulting when under-statisticized studies are excluded from meta-analyses. *Journal of Educational Measurement, 24*(1), 72-76.

STOCK, W.A., OKUN, M.A., HARING, M.J., MILLER, W. & KINNEY,C. & CEURVORST, R.W. (1982). Rigor in data synthesis: A case study of reliability in meta-analysis. *Educational Researcher, 11* (6), 10-14.

STOCK, W.A., OKUN, M.A., HARING, M.J. & WITTER, R.A. (1983). Age differences in subjective well-being. In R.J. Light (Ed.), *Evaluation Studies Review Annual* (Vol.8, pp.279-302). Beverly Hills: Sage Publications.

STRAUSS, S. (1969). Guidelines for analysis of research reports. *Journal of Educational Research, 63* (4), 165-169.

STRIKE, K. & POSNER, G. (1983). Types of synthesis and their criteria. In S.A. Ward & L.J.Reed (Eds.), *Knowledge structure and use: Implications for synthesis and interpretation* (pp.343-362). Philadelphia, PA: Temple University Press.

STRINGFIELD, J.K. (1991). The Humpty Dumpty school of communications in education. *The Educational Forum, 55 (3)*, 261-270.

STRUBE, M.J., GARDNER, W. & HARTMANN, D.P. (1985). Limitations, liabilities, and obstacles in reviews of the literature: The current status of meta-analysis. *Clinical Psychology Review*, *5*, 63-78.

STRUBE, M.J. & HARTMANN, D.P. (1982). A critical appraisal of meta-analysis. *British Journal of Clinical Psychology*, *21*, 129-139.

STRUBE, M.J. & HARTMANN, D.P. (1983). Meta-analysis: Techniques, applications and function. *Journal of Consulting and Clinical Psychology*, *51*, 14-27.

STRUBE, M.J. & MILLER, R.H. (1986). Comparison of power rates for combined probability procedures: A simulation study. *Psychological Bulletin*, *99* (3), 407-415.

TEDESCHI, J.T., GAES, G.G., RIORDAN, C. & QUIGLEY-FERNANDEZ, B. (1981). Social psychology and cumulative knowledge. *Personality and Social Psychology Bulletin*, *7*, 161-172.

WALBERG, H.J. (1984). Quantification reconsidered. In E. Gordon (Ed.), *Review of Research in Education* (pp.369-402). Washington, D.C.: American Educational Research Association.

WALBERG, H.J. (1984b). Improving the productivity of America's schools. *Educational Leadership*, *41* (8), 19-27.

WALBERG, H.J. (1986). Synthesis of research on teaching. In M.C. Wittrock (Ed.), *Handbook of research on teaching* (Third edition, pp.214-229). Washington, D.C.: American Educational Research Association.

WALBERG, H.J., SCHILLER, B. & HAERTEL, G.D. (1979). The quiet revolution in educational research. *Phi Delta Kappan*, *61*, 179-183.

WANOUS, J.P., SULLIVAN, S.E. & MALINAK, J. (1989). The role of judgment calls in meta-analysis. *Journal of Applied Psychology*, *74* (2), 259-264.

WARD, S.A. & REED, L.J. (1983). Introduction. In S.A. Ward & L.J. Reed (Eds.), *Knowledge structure and use: Implications for synthesis and interpretation* (pp.1-17). Philadelphia, PA: Temple University Press.

WAXMAN, H. & WALBERG, H.J. (1982). The relation of teaching and learning. A review of reviews of process-product research. *Contemporary Education Review*, *1*, 103-120.

WITTER, R.A., OKUN, M.A., STOCK, W.A. & HARING, M.J. (1984). Education and subjective well-being: A meta-analysis. *Educational Evaluation and Policy Analysis*, *6* (2), 165-173.

WORTMAN, P.M. (1983). Meta-analysis: A validity perspective. In R.J. Light (Ed.), *Evaluation Studies Review Annual* (Vol.8, pp.157-166). Beverly Hills: Sage Publications.

WORTMAN, P.M. (1983b). Evaluation research: A methodological perspective. *Annual Review of Psychology*, *34*, 223-260.

WORTMAN, P.M. & BRYANT, F.B. (1985). School desegration and black achievement. An integrative review. *Sociological Methods & Research*, *13* (3), 289-324.

APPENDIX

1.	Meta-analyses, replications and critiques	111
2.	Coding Tables	
	Studies 1 to 12	116
	Studies 13 to 22	119
	Studies 23 to 31	122
	Studies 32 to 40	125
	Studies 41 to 48	128
3. A.	Frequency of recordings per year	131
3. B.	Percentage of recordings per year	134

1. Meta-analyses, replications and critiques
(indented sections refer to replications or critiques)

1 ABRAMI, P.C., LEVENTHAL, L. & PERRY, R.P. (1982). Educational seduction. *Review of Educational Research*, *52* (3), 446-464.

2 COHEN, P.A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research*, *51* (3), 281-309.

3 COHEN, P.A., KULIK, J.A. & KULIK, C-L.C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal*, *19* (2), 237-248.

4 DUSEK, J.B. & JOSEPH, G. (1983). The bases of teacher expectancies: A meta-analysis. *Journal of Educational Psychology*, *75* (3), 327-346.

5 FISHER, C.D. & GITELSON, R. (1983). A meta-analysis of the correlates of role conflict and ambiguity. *Journal of Applied Psychology*, *68* (2), 320-333.

6a FRICKE, R. (1985). Klarheit der Lehrersprache und Schülerleistung: Eine Meta-analyse. In K. Aurin & B. Schwarz (Hrsg.), *Die Erforschung pädagogischer Wirkungsfelder* (S.205-211). Freiburg: Arbeitsgruppe für empirische pädagogische Forschung in der DGfE.
6b FRICKE, R. & TREINIES, G. (1985). *Einführung in die Metaanalyse*. Bern: Verlag Hans Huber.

7a GLASS, G.V. & SMITH, M.L. (1979). Meta-analysis of research on class size and achievement. *Educational Evaluation and Policy Analysis*, *1* (1), 2-16.
7b SMITH, M.L. & GLASS, G.V. (1980). Meta-analysis of research on class size and its relationship to attitudes and instruction. *American Educational Research Journal*, *17* (4), 419-433.
- CAHEN, L.S. & FILBY, N.N. (1979). The class size achievement issue: New evidence and a research plan. *Phi Delta Kappan*, *60*, 492-496.
- EDUCATIONAL RESEARCH SERVICE (1980). Class size research: A critique of recent meta-analysis. *Phi Delta Kappan*, *62*, 239-241.
- GLASS, G.V., CAHEN, L.S., SMITH, M.L. & FILBY, N.N. (1982). *School class size: Research and policy*. Beverly Hills: Sage Publications.
- HEDGES, L.V. & STOCK, W. (1983). The effect of class size: An examination of rival hypotheses. *American Educational Research Journal*, *20* (1), 63-85.
- JACKSON, G.A. (1983). School class size: Research and policy (Book review). *Harvard Educational Review*, *53* (1), 74-77.
- PREECE, P.F.W. (1987). Class size and learning: A theoretical model. *Journal of Educational Research*, *80* (6), 377-379.
- SIMPSON, S.N. (1980). Comments on „Meta-analysis of research on class size achievement". *Educational Evaluation and Policy Analysis*, *3*, 81-83.
- SLAVIN, R.E. (1984). Meta-analysis in education: How has it been used? *Educational Researcher*, *13* (8), 6-15.

8 GRAUE, M.E., WEINSTEIN, T. & WALBERG, H.J. (1983). School-based home instruction and learning: A quantitative synthesis. *Journal of Educational Research*, *76* (6), 351-360.

9 IDE, J.K., PARKERSON, J.A., HAERTEL, G.D. & WALBERG, H.J. (1981). Peer group influence on educational outcomes: A quantitative synthesis. *Journal of Educational Psychology*, *73*(4), 472-484.

10 IVERSON, B,K. & WALBERG, H.J. (1982). Home environment and school learning: A quantitative synthesis. *Journal of Experimental Education*, *50*, 144-151.

11 JACOBS, B. (1987). Die Auswirkungen transparenzschaffender Maßnahmen auf die aktuelle Angst der Schüler vor einer Klassenarbeit – Eine Metaanalyse zum Saarbrücker Schulangstprojekt. *Empirische Pädagogik*, *1* (2), 139-160.

12 JOHNSON, D.W., MARUYAMA, G., JOHNSON, R., NELSON, D. & SKON, L. (1981). Effects of cooperative, competitive and individualistic goal structures on achievement: A meta-analysis. *Psychological Bulletin*, *89* (1), 47-62.

– COTTON, J.L. & COOK, M.S. (1982). Meta-analysis and the effect of various reward systems: Some different conclusions from Johnson et al. *Psychological Bulletin*, *92*, 176-183.

– JOHNSON, D.W., MARUYAMA, G., JOHNSON, R.T. (1982). Separating ideology from currently available data. A reply to Cotton, Cook and McGlynn. *Psychological Bulletin*, *92*, 186-192.

– McGLYNN, R.P. (1982). A comment on the meta-analysis of goal structures. *Psychological Bulletin* , *92*, 184-185.

– SLAVIN, R.E. (1984). Meta-analysis in education: How has it been used? *Educational Researcher* , *13* (8), 6-15.

13a KLAUER, K.J. (1981). Zielorientiertes Lehren und Lernen bei Lehrtexten. *Unterrichtswissenschaft*, *4*, 300-318.

13b KLAUER, K.J. (1984). Intentional and incidental learning with instructional texts: A meta-analysis for 1970 – 1980. *American Educational Research Journal*, *21* (2), 323-339.

– HAGER, W. (1985). Beurteilung inhaltlicher Hypothesen als Alternative zur Metaanalyse: Wirken Zielvorgaben aufmerksamkeitslenkend oder allgemein motivierend? *Psychologische Beiträge*, *27*, 200-217.

14 KLINZING, H.G. (1984). Expressives nichtverbales Lehrerverhalten im Unterricht: Ein Forschungsbericht. *Unterrichtswissenschaft*, *4*, 308-309.

15 KLINZING, H.G., TISHER, R.P. & KLINZING-EURICH, G. (1984). Training nichtverbaler Wahrnehmungs- und Ausdrucksfähigkeit. Befunde und Empfehlungen für die Entwicklung von Trainingskursen. *Unterrichtswissenschaft*, *4*, 320-339.

16 KREMER, B.K. & WALBERG, H.J. (1981). A synthesis of social and psychological influences on science learning. *Science Education*, *65*, 11-23.

17 KULIK, J.A., KULIK, C.-L.C. & COHEN, P.A. (1979). A meta-analysis of outcome studies of Keller's personalized system of instruction. *American Psychologist, 34*, 307-318.

18 LYSAKOWSKI, R.S. & WALBERG, H.J. (1981). Classroom reinforcement and learning: A quantitative synthesis. *Journal of Educational Research, 75* (2), 69-77.

19 LYSAKOWSKI, R.S. & WALBERG, H.J. (1982). Instructional effects of cues, participation and corrective feedback: A quantitative synthesis. *American Educational Research Journal, 19* (4), 559-578.
– SLAVIN, R.E. (1984). Meta-analysis in education: How has it been used? *Educational Researcher , 13* (8), 6-15.

20 PASCHAL, R.A., WEINSTEIN, T. & WALBERG, H.J. (1984). The effects of homework on learning: A quantitative synthesis. *Journal of Educational Research, 78*, 97-104.

21 PFLAUM, S.W., WALBERG, H.J, KAREGIANES, M.L. & RASHER, S. (1980). Reading instruction: A quantitative synthesis. *Educational Researcher, 9*, 12-18.
– SLAVIN, R.E. (1984). Meta-analysis in education: How has it been used? *Educational Researcher, 13* (8), 6-15.

22a RAUDENBUSH, S.W. (1983). Utilizing controversy as a source of hypotheses for meta-analysis. In R.J. Light (Ed.), *Evaluation Studies Review Annual* (Vol.8, pp.303-325). Beverly Hills: Sage Publications.
22b RAUDENBUSH, S.W. (1984). Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy induction: A synthesis of findings from 18 experiments. *Journal of Educational Psychology, 76* (1), 85-97.

23 REDFIELD, D.L. & ROUSSEAU, E.W. (1981). A meta-analysis of experimental research on teacher questioning behavior. *Review of Educational Research, 51* (2), 237-245.

24 SAMSON, G.E., STRYKOWSKI, B., WALBERG, H.J. & WEINSTEIN, T. (1987). The effects of teacher questioning levels on student achievement: A quantitative synthesis. *Journal of Educational Research, 80* (5), 290-295.

25a STOCK, W.A., OKUN, M.A., HARING, M.J., MILLER, W. & KINNEY,C. & CEURVORST, R.W. (1982). Rigor in data synthesis: A case study of reliability in meta-analysis. *Educational Researcher, 11* (6), 10-14.
25b STOCK, W.A., OKUN, M.A., HARING, M.J. & WITTER, R.A. (1983). Age differences in subjective well-being. In R.J. Light (Ed.), *Evaluation Studies Review Annual* (Vol.8, pp.279-302). Beverly Hills: Sage Publications.
25c WITTER, R.A., OKUN, M.A., STOCK, W.A. & HARING, M.J. (1984). Education and subjective well-being: A meta-analysis. *Educational Evaluation and Policy Analysis, 6* (2), 165-173.

26 UGUROGLU, M.E. & WALBERG, H.J. (1979). Motivation and achievement: A quantitative synthesis. *American Educational Research Journal, 16* (4), 375-389.

27 WAXMAN, H.C., WANG, M.C., ANDERSON, K.A. & WALBERG, H.J. (1985). Adaptive education and student outcomes: A quantitative synthesis. *Journal of Educational Research*, *78* (4), 228-236.

28a WORTMAN, P.M. & BRYANT, F.B. (1985). School desegration and black achievement. An integrative review. *Sociological Methods & Research*, *13* (3), 289-324.
28b BRYANT, F.B. & WORTMAN, P.M. (1984). Methodological issues in the meta-analysis of quasi-experiments. In W.H. Yeaton & P.M. Wortman (Eds.), *Issues in Data Synthesis*. New Directions for Program Evaluation, (no.24, pp.5-24). San Francisco: Jossey-Bassey.

29 LUITEN, J., AMES, W. & ACKERSON, G. (1980). A meta-analysis of the effects of advance organizers on learning and retention. *American Educational Research Journal*, *17* (2), 211-218.

30 STONE, C.L. (1983). A meta-analysis of advance organizer studies. *Journal of Experimental Education*, *51*, 194-149.

31 MOORE, D.W. & READENCE, J.E. (1984). A quantitative and qualitative review of graphic advance organizer research. *Journal of Educational Research*, *78* (1), 11-17.

32 HANSFORD, B.C. & HATTIE, J.A. (1982). The relationship between self and achievement / performance measures. *Journal of Educational Research*, *52* (1), 123-142.
– HATTIE, J.A. & HANSFORD, B.C. (1982). Selfmeasures and achievement. Comparing a traditional review of literature with a meta-analysis. *Australian Journal of Education* , *26*, 71-75.

33 KULIK, J.A., COHEN, P.A. & EBELING, B.J. (1980). Effectiveness of programmed instruction in higher education: A meta-analysis of findings. *Educational Evaluation and Policy Analysis*, *2* (6), 51-64.

34 COHEN, P.A., EBELING, B.J. & KULIK, J.A. (1981). A meta-analysis of outcome studies of visual based instruction. *Educational Communication and Technology Journal*, *29*, 26-36.

35 KULIK, C.C. & KULIK, J.A. (1982). Effects of ability grouping on secondary school students: Meta-analysis of evaluation findings. *American Educational Research Journal*, *19*, 415-428.
– SLAVIN, R.E. (1984). Meta-analysis in education: How has it been used? *Educational Researcher* , *13* (8), 6-15.

36 KULIK, C.-L.C., SCHWALB, B.J. & KULIK, J.A. (1982). Programmed instruction in secondary education: A meta-analysis of evaluation findings. *Journal of Educational Research*, *75* (3), 133-138.

37 BANGERT, R.L., KULIK, J.A. & KULIK, C-L.C. (1983). Individualized systems of instruction in secondary schools. *Review of Educational Research*, *53* (2), 143-158.

38 BANGERT-DROWNS, R.L., KULIK, J A. & KULIK, C-L.C. (1983). Effects of coaching programs on achievement test performance. *Review of Educational Research*, *53* (4), 571-585.

39 KULIK, J.A. & KULIK, C.-L.C. (1984). Effects of accelerated instruction on students. *Review of Educational Research*, *54* (3), 409-425.

40 KULIK, J.A. & KULIK, C.-L.C. (1988). Timing of feedback and verbal learning. *Review of Educational Research*, *58* (1), 79-97.

41 BOULANGER, F.D. (1981). Ability and science learning: A quantitative synthesis. *Journal of Research in Science Teaching*, *18* (4), 113-121.

42 BREDDERMAN, T. (1983). The effects of activity-based elementary science on student outcomes: A quantitative synthesis. *Review of Educational Research*, *53*, 499-518.

43 DEKKERS, J. & DONATTI, S. (1981). The integration of research studies on the use of simulation as an instructional strategy. *Journal of Educational Research*, *74*, 424-427.

44 DRUVA, C.A. & ANDERSON, R.D. (1983). Science teacher characteristics by teacher behavior and by student outcome: A meta-analysis of research. *Journal of Research in Science Teaching*, *20*, 467-479.

45 FINDLEY, M.J. & COOPER, H.M. (1983). Locus of control and academic achievement: A literature review. *Journal of Personality and Social Psychology*, *44*, 419-427.

46 GIACONIA, R.M. & HEDGES, L.V. (1982). Identifying features of effective open education. *Review of Educational Research*, *52* (4), 579-602.

47 MABE III, P.A. & WEST, S.G. (1982). Validity of self evaluation of ability. *Journal of Applied Psychology*, *67*, 280-296.

48 STEINKAMP, M.W. & MAEHR, M.L. (1983). Affect, ability and science achievement: A quantitative synthesis of correlational research. *Review of Educational Research*, *53* (3), 369-396.

2. Coding Tables

(Parenthesis refer to uncertain or partially fitting codings)

Study	1	2	3	4	5	6a	6b	7a	7b	8	9	10	11	12
I Theoretical framework														
1 -problem orientated	x	x		x	x	x	x	x		x	x	x	x	x
-indirectly deducible			x						x					
2 -prior reviews mentioned	x	x	x	x	x	x	x	x	x	x	x	x		x
-prior research mentioned	x	x		x	x	x	x	x		x	x	x	x	
-shortcomings indicated	x	x	x	x	x	x	x	x	x	x				x
3 Construct definitions														
-specifically formulated	x	x		x		x	x	x			x	x	x	x
-indirectly deducible			x		x			x	x	x	x	x		x
-diversity ignored					x			x						
-diversity analyzed	x	x	x			x	x		x	x	x	x	x	x
4 Aims														
-exploratory	x	x	x	x	x	x	x	x		x	x	x		x
-specific hypotheses				x									x	
II Sampling														
1 Search strategy														
-retrieval system		x	x	x	x			x		x	x	x		
-bibliographies		x	x			x	x	x		x		x		
-experts quizzed														
2 Representativeness														
-published/unpublished		x	x		x			x	x	x				
-file drawer/fail safe				x	x		x	(x)						x
-source variation		x	x					x	x	x				x
3 Selection criteria														
-specified		x	x	x						x	x	x		x
-excluded studies				x				x			x			
-articles listed	x	x		x	x		x			x	x	x		x
-available on request			x					x	x					
4 Number of studies	9	41	65	77	43	9	9	77	59	29	10	18	9	122

Study continued	1	2	3	4	5	6a	6b	7a	7b	8	9	10	11	12
III Coded characteristics														
1 sample size	x				x		x	(x)			x	x	x	
2 person characteristics														
-sex				x	x						x	x		x
-age/class/grade			x		x	x	x	x	x	x	x	x	x	x
-educational variables		x	x	x	x			x		x		x		
-demographic variables				x						x	x	x		
3 settings														
-laboratory	x			x		(x)	x			x				x
-field		x	x		x	(x)	x			x	x	x	x	x
-region							x	x		x	x	x		
-publication date	x	x	x	x				x	x	x	x	x		x
4 study features														
-global quality index		x					x			x			x	(x)
-sampling/assignment		x	x					x	x	x				
-operationalizations	x	x	x		x		x			x	x	x	x	x
-reliability of outcome		x	x		x		x	x		x	x			
-interaction/control var.	x	x	x								x			
-unit of analysis		x								x			x	
-adequacy of stat.anal.											x			
IV Data analysis														
1 Effect size conceptualization														
-formula/index indicated	x	x	x	x	x		x	x	x	(x)	x	(x)	x	x
-magnitudes listed per study	x						x					x		
-stem-leaf/graph summary		x	x					x			x		x	
-significance indicated		x	x	x			x							
-outcome variable indicated	x	x	x	x	x		x				x	x	x	
2 Coding quality														
-missing data problem noted				x	x			x	x		x			x
-information gained elsewhere	x													
-studies excluded			x								x			
-global intercoder agreement		x												x
-indices per item														
-strategy to reach agreement										x				
3 Unit of analysis														
-study	(x)		x	x			x						x	
-study findings		x			x			x	x	x	x	x		x
4 Non-independence														
-present,taken into account			x	x		x	x			x	x	x	(x)	
-ignored, authors	x	x		x	x		x					x	(x)	x
-ignored, findings	(x)	x			x			x	x					x

Study continued	1	2	3	4	5	6a	6b	7a	7b	8	9	10	11	12
5 Average effect sizes for														
-total sample	x						x	x	x	x	x	x		
-subsample		x	x	x	x		x	x	x	x	x	x	x	x
-range/std.dev./std.error	x	x	x		x		x	x	x	x	x	x	x	x
6 Statistical analysis														
-conventional		x						x	x	x	x	x		x
-'modern'				x	x		x						x	x
-assumptions tested														
-limitations noted	x						x	x	x		x	x		x
-parallel analysis							x			x	x	x		x
-capitalization on chance														
7 Effect size variability														
-heterogeneity tested					x		x			x			x	
-quality aspects		x	x					x	x	x	x			x
-treatment variation	x	x	x	(x)		x	x			x	x	x	x	x
-outcome variation		x	x	(x)	x			x	x	x	x	x	x	x
-subject characteristics	x	x	x	(x)	x			x		x	x	x		x
-design/stat.analysis		x									x			x
-contextual/scope variables		x	x					x	x	x	x	x		x
V Interpretation														
1 Effect size														
-std.dev/percentile	x		x	x			x	x	x	x	x		x	x
-binomial effect size display														
-Cohen's classification	x	x	x	x										
-behavioral indices														
-expert judgement														
2 Theoretical implications														
-old theory/impressions		x	x	(x)	x	x				x			x	
-new theory/hypotheses														x
3 Practical implications														
-for policy or practice	x	x		(x)			x							x
-limitations noted	x	x					x	x					x	x
4 Future implications														
-for primary research	x	x		x	x			x		x	x	x	x	x
-for reviews				x			x							

Study	13a	13b	14	15	16	17	18	19	20	21	22a	22b
I Theoretical framework												
1 -problem orientated	x	x	x			x	x	x		x	x	x
-indirectly deducible				x	x				x			
2 -prior reviews mentioned	x	x	x		x	x	x	x	x	x	x	x
-prior research mentioned	x		x	x		x	x	x	x	x	x	x
-shortcomings indicated		x									x	x
3 Construct definitions												
-specifically formulated	x	x	x	x	x	x	x	x	x	x		
-indirectly deducible			x	x			x		x	x	x	x
-diversity ignored			x									
-diversity analyzed							x	x	x	x		
4 Aims												
-exploratory	x	x	x	x	x	x	x	x	x	x	(x)	x
-specific hypotheses	x	x									x	x
II Sampling												
1 Search strategy												
-retrieval system	x	x		x	x	x	x	x	x		x	x
-bibliographies		x		x	x	x	x	x		x	x	x
-experts quizzed	x	x										
2 Representativeness												
-published/unpublished					x			x	x			x
-file drawer/fail safe										(x)		
-source variation									x			x
3 Selection criteria												
-specified	x	x			x	x	x	x	x		x	x
-excluded studies	x				x		x	x		(x)		x
-articles listed	x	x	x	(x)	x		x	x	x			x
-available on request				(x)		x				x		
4 Number of studies	16	24	(35)	(64)	20	72	39	(54)	15	97	18	18
III Coded characteristics												
1 sample size	x	x	x	(x)	x		(x)	(x)				
2 person characteristics												
-sex				(x)	(x)		x	x		x	x	
-age/class/grade	x		x	(x)	x		x	x	x	x	x	x
-educational variables			x			x	x		x	x		
-demographic variables			x		x		x	x	x	x	x	

Study continued	13a	13b	14	15	16	17	18	19	20	21	22a	22b
3 settings												
-laboratory	(x)		x	x								
-field			x	x	x	x	(x)	(x)				x
-region					x		x	x	x			
-publication date	x	x	x	x	x		x	x	x			x
4 study features												
-global quality index				(x)	x							x
-sampling/assignment						x			x	x		x
-operationalizations			x	x	x	x	x	x	x	x	x	x
-reliability of outcome	x	x				x	x	x	x	x	(x)	
-interaction/control var.						x	x	x		x		
-unit of analysis									x			
-adequacy of stat.anal.							x	x	x			x
IV Data analysis												
1 Effect size conceptualization												
-formula/index indicated	x	x		x	x	x	x	x	x	x	x	x
-magnitudes listed per study	x	x	(x)	x	x							x
-stem-leaf/graph summary					x	x		x	x	x		x
-significance indicated			x	x								x
-outcome variable indicated	x	x	x	x	x	x						x
2 Coding quality												
-missing data problem noted	x	x		x	x		x	x		x	x	
-information gained elsewhere												
-studies excluded					x		x	x	x	x		
-global intercoder agreement							x	x				
-indices per item												
-strategy to reach agreement										x		
3 Unit of analysis												
-study						(x)						x
-study findings	x	x			x		x	x	x	x	x	x
4 Non-independence												
-present,taken into account						x	x	x	x	x		
-ignored, authors	x	x	x	x	x		x	x	x			x
-ignored, findings	x	x	x	x	x				(x)	(x)	x	x
5 Average effect sizes for												
-total sample			(x)	x			x	x	x	x	x	x
-subsample	x	x	x	x	x	x	x	x	x	x	x	x
-range/std.dev./std.error	x	x				(x)	x	x	x	x	x	x

Study continued	13a	13b	14	15	16	17	18	19	20	21	22a	22b
6 Statistical analysis												
-conventional	x	x	(x)	(x)	(x)	x	x	x	x	x		
-'modern'											x	x
-assumptions tested												
-limitations noted	x	x									(x)	x
-parallel analysis	x	x					x				x	x
-capitalization on chance	x	x						x		x		
7 Effect size variability												
-heterogeneity tested							x	x	x			x
-quality aspects				(x)		x	x	x	x	x	x	x
-treatment variation	x	x	x	x	(x)	x	x	x	x	x	x	x
-outcome variation	x	x	x	x	(x)	x		x	x	x		
-subject characteristics						x	x	x	x	x	x	x
-design/stat.analysis				x		x			x	x		
-contextual/scope variables				x		x	x	x	x	x	x	x
V Interpretation												
1 Effect size												
-std.dev/percentile						x	x	x		x	x	x
-binomial effect size display												x
-Cohen's classification						x						
-behavioral indices												
-expert judgement												
2 Theoretical implications												
-old theory/impressions	x	x	x		(x)	x	x	x		x	x	x
-new theory/hypotheses											(x)	
3 Practical implications												
-for policy or practice	x	x	x	x						x		
-limitations noted	x	x	x	x		x	x	x		x	x	x
4 Future implications												
-for primary research				(x)	x			x	x	x	x	x
-for reviews						(x)				x	x	x

Study	23	24	25a	25b	25c	26	27	28a	28b	29	30	31
I Theoretical framework												
1 -problem orientated	x	x		x	x			x			x	x
-indirectly deducible						x	x		x	x		
2 -prior reviews mentioned	x	x		x		x	x	x		x	x	x
-prior research mentioned		x		x	x		x	x		x		x
-shortcomings indicated				x		x		x		x		x
3 Construct definitions												
-specifically formulated	x	x		x	x		x		x	x	x	x
-indirectly deducible			x	x	x	x	x	x		x	x	
-diversity ignored												
-diversity analyzed				x	x	x	x		(x)			x
4 Aims												
-exploratory	x	x		x		x	x	x	x			
-specific hypotheses					x							
II Sampling												
1 Search strategy												
-retrieval system		x		x		x	x	x			x	x
-bibliographies	x	x		x		x	x	x			x	x
-experts quizzed				x	x			x				
2 Representativeness												
-published/unpublished				x	x		x	x	x	x		x
-file drawer/fail safe												
-source variation				x			x	x	x			x
3 Selection criter												
-specified	x	x		x		x	x					x
-excluded studies	x			(x)	x		x	x	x		x	
-articles listed	x	x				(x)	x	(x)			x	
-available on request				x	x					x		
4 Number of studies	14	14		(147)	(116)	40	38	31	31	135	29	23
III Coded characteristics												
1 sample size						(x)	(x)					
2 person characteristics												
-sex			x	x	x	x	x				x	
-age/class/grade		x	x	x	x	x	x	x		x	x	x
-educational variables		x	x	x	x		x			x	x	
-demographic variables		x	x	x	x		x					

Study continued	23	24	25a	25b	25c	26	27	28a	28b	29	30	31
3 settings												
-laboratory	x	x										
-field	x	x		x			x	x	x	(x)		
-region				x	x	x	x		x			
-publication date			x	x	x	x	x	x	x			
4 study features												
-global quality index	x		x	x	x			x				
-sampling/assignment		x	x	x	x		x	x	x			
-operationalizations	x	x	x	x	x	x	x	x	x	x	x	x
-reliability of outcome		x		x		x	x		x			
-interaction/control var.				x	x		x		x			
-unit of analysis							x					
-adequacy of stat.anal.		x			x		x	x	x			
IV Data analysis												
1 Effect size conceptualization												
-formula/index indicated	x	x		x	x	x	x	x	x	x	x	x
-magnitudes listed per study		x					x	(x)				
-stem-leaf/graph summary				x	x	x	x					
-significance indicated												
-outcome variable indicated		x					x	x				
2 Coding quality												
-missing data problem noted	x	x		x	x	x	x	x	x	x	x	x
-information gained elsewhere	x			(x)			x	x	x			
-studies excluded	x	x		x	x		x	x	x		x	
-global intercoder agreement		x			(x)		x					x
-indices per item			x									
-strategy to reach agreement			x	x				x				
3 Unit of analysis												
-study	(x)							x				
-study findings	(x)	x		x	x	x	x	x		x	x	x
4 Non-independence												
-present,taken into account	(x)	x		x	x	x	x	x				
-ignored, authors	x	x					x				x	
-ignored, findings		x		x		x				(x)	x	x
5 Average effect sizes for												
-total sample	x	x		(x)	(x)	x	x	x			x	x
-subsample	x	x		x	x	x	x	x		x	x	x
-range/std.dev./std.error		x		x	x	x	x	x		x	x	x

Study continued	23	24	25a	25b	25c	26	27	28a	28b	29	30	31
6 Statistical analysis												
-conventional		x		x	x	x	x	x				
-'modern'				x	x							
-assumptions tested				x	x			x				
-limitations noted				x				x				
-parallel analysis				x	x							
-capitalization on chance												
7 Effect size variability												
-heterogeneity tested												
-quality aspects	x	x		x	x		x	x				
-treatment variation	x	x					x			x	x	x
-outcome variation		x		x	x	x	x	x		x	x	x
-subject characteristics		x		x	x	x	x	x		x	x	x
-design/stat.analysis		x		x		x	x	x				
-contextual/scope variables	x	x		x	x	x	x	x		x	x	x
V Interpretation												
1 Effect size												
-std.dev/percentile	x					x	x	x		x	x	x
-binomial effect size display												
-Cohen's classification												x
-behavioral indices								(x)				
-expert judgement								(x)				
2 Theoretical implications												
-old theory/impressions	x	x		x	x	x	(x)	x		x	x	x
-new theory/hypotheses												
3 Practical implications												
-for policy or practice	x							x				
-limitations noted	x			x	x		(x)	x	x	x		x
4 Future implications												
-for primary research				x		x				x		x
-for reviews	x		x					x	x			

Study	32	33	34	35	36	37	38	39	40
I Theoretical framework									
1 -problem orientated	x	x	x	x			x	x	x
-indirectly deducible					x	x			
2 -prior reviews mentioned	x	x	x	x	x	x	x	x	x
-prior research mentioned		x		x			x		x
-shortcomings indicated		x	x	x	x	x	(x)	x	(x)
3 Construct definitions									
-specifically formulated	x		x			x	x	x	
-indirectly deducible	x	x		x	x	x			x
-diversity ignored									
-diversity analyzed	x	x	x	x	x	x	x	x	x
4 Aims									
-exploratory	x	x	x	x	x	(x)	x	x	x
-specific hypotheses									
II Sampling									
1 Search strategy									
-retrieval system	x	x	x	x	x	x	x	x	x
-bibliographies		x	x	x	x	x	x	x	x
-experts quizzed									
2 Representativeness									
-published/unpublished	x	x	x	x	x	x	x	x	x
-file drawer/fail safe									(x)
-source variation	x	x	x	x	x	x	x	x	x
3 Selection criteria									
-specified		x	x	x	x	x	x	x	x
-excluded studies								(x)	
-articles listed						(x)	x	x	x
-available on request	x	x	x	x	x				
4 Number of studies	128	56	72	52	48	(51)	(25)	21	40
III Coded characteristics									
1 sample size	(x)								
2 person characteristics									
-sex	x								
-age/class/grade	x		(x)	x	x	x	x	x	x
-educational variables	x		(x)	x	x		x		
-demographic variables	x								

Study continued	32	33	34	35	36	37	38	39	40
3 settings									
-laboratory							x		x
-field		x	x		x	x	x		x
-region									
-publication date	x	x	x	x	x	x	x	x	x
4 study features									
-global quality index	x								
-sampling/assignment		x	x	x	x	x	x	x	x
-operationalizations	x	x	x	x	x	x	x	x	x
-reliability of outcome	x	x	(x)	(x)	(x)		(x)		
-interaction/control var.		x	x	x	x	x		(x)	
-unit of analysis									
-adequacy of stat.anal.									
IV Data analysis									
1 Effect size conceptualization									
-formula/index indicated	x	x	x	x	x	x	x	x	x
-magnitudes listed per study						(x)	x	(x)	x
-stem-leaf/graph summary	x	(x)	(x)	(x)	(x)		(x)	(x)	
-significance indicated		(x)	(x)	(x)		(x)			
-outcome variable indicated		(x)	(x)	(x)	(x)	(x)	x	(x)	(x)
2 Coding quality									
-missing data problem noted	x	x		x	x				
-information gained elsewhere									
-studies excluded	x								
-global intercoder agreement			x						
-indices per item									
-strategy to reach agreement	x	x	x						x
3 Unit of analysis									
-study		x	x	x	x	x	x	x	x
-study findings	x								
4 Non-independence									
-present,taken into account	x	x	x	x	x	(x)	x	x	x
-ignored, authors						x	x	x	x
-ignored, findings	x		(x)						
5 Average effect sizes for									
-total sample	x						x		
-subsample	x	x	x	x	x	x	x	x	x
-range/std.dev./std.error	x	(x)	(x)	x	x	x	x	x	(x)

Study continued	32	33	34	35	36	37	38	39	40
6 Statistical analysis									
-conventional	x	x	x	x	x	x	x	(x)	x
-'modern'							x		
-assumptions tested									
-limitations noted	x								
-parallel analysis	x	(x)							
-capitalization on chance									
7 Effect size variability									
-heterogeneity tested							x		
-quality aspects	x								
-treatment variation	(x)	x	x	x	x	x	x	x	x
-outcome variation	(x)	x	x	x	(x)	(x)	(x)	(x)	(x)
-subject characteristics	x			x	x	x	x	x	x
-design/stat.analysis		x	x	x	x	x	x	x	
-contextual/scope variables	x	x	x	x	x	x	x	x	x
V Interpretation									
1 Effect size									
-std.dev/percentile	x	x	x	x	x	x	x	x	x
-binomial effect size display									
-Cohen's classification		x	x	x	x	x	x	x	
-behavioral indices		x	x				x	x	
-expert judgement									
2 Theoretical implications									
-old theory/impressions	x	x	x	x	x	x	x	x	x
-new theory/hypotheses									
3 Practical implications									
-for policy or practice						(x)			x
-limitations noted	x	x		x	x			x	x
4 Future implications									
-for primary research	x		x		x	x			x
-for reviews	x								

Study	41	42	43	44	45	46	47	48
I Theoretical framework								
1 -problem orientated	x	x	x		x	x	x	x
-indirectly deducible				x			x	x
2 -prior reviews mentioned	x	x			x	x		x
-prior research mentioned	x		x		x		x	x
-shortcomings indicated			x		x	x		
3 Construct definitions								
-specifically formulated	x	x		x	x	x	x	x
-indirectly deducible		x	x				x	
-diversity ignored								
-diversity analyzed	x	x		x	x	x		(x)
4 Aims								
-exploratory	x	x	x	x	x	x	x	x
-specific hypotheses							(x)	
II Sampling								
1 Search strategy								
-retrieval system		x	x	(x)	x		(x)	x
-bibliographies		x	x			(x)		x
-experts quizzed								
2 Representativeness								
-published/unpublished		x	x	x	x			
-file drawer/fail safe					x		(x)	
-source variation		x	x		(x)			
3 Selection criteria								
-specified	x		x	(x)	x		x	x
-excluded studies								
-articles listed	x	x					x	x
-available on request				x				
4 Number of studies	34	57	93	65	98	153	(52)	66
III Coded characteristics								
1 sample size		(x)			(x)		x	x
2 person characteristics								
-sex		x		x	x	x		x
-age/class/grade	x	x	x	x	x	x	x	x
-educational variables	x	x		x				
-demographic variables	x	x			x	x		(x)

Study continued	41	42	43	44	45	46	47	
3 settings								
-laboratory								
-field	(x)	(x)						x
-region					x			(x)
-publication date	x	x	x		x	x	x	x
4 study features								
-global quality inde						x		x
-sampling/assignment		x				(x)		
-operationalizations	x	x	(x)	x	x	x	x	x
-reliability of outcome	x	(x)	(x)				(x)	x
-interaction/control var.		x						x
-unit of analysis	x							
-adequacy of stat.anal.	x							
IV Data analysis								
1 Effect size conceptualization								
-formula/index indicated	x	x	x	x	x	x	x	x
-magnitudes listed per study							x	x
-stem-leaf/graph summary							x	x
-significance indicated					(x)			
-outcome variable indicated		x					x	x
2 Coding quality								
-missing data problem noted	x	x	x		x	x	x	x
-information gained elsewhere		(x)					x	x
-studies excluded		x	x		(x)	x		
-global intercoder agreement	x						x	x
-indices per item							(x)	
-strategy to reach agreement								x
3 Unit of analysis								
-study			(x)		(x)			
-study findings	x	(x)		x	(x)	x	x	x
4 Non-independence								
-present,taken into account	(x)	x			(x)	x		(x)
-ignored, authors	x	x					x	x
-ignored, findings	x			x	(x)		x	x
5 Average effect sizes for								
-total sample		x	x		x		x	
-subsample	x	x	x	x	x	x	x	x
-range/std.dev./std.error	x	x	x	x	x	x	x	x

Study continued	41	42	43	44	45	46	47	48
6 Statistical analysis								
-conventional	x	x	x	x	x		x	x
-'modern'					x	x	x	x
-assumptions tested					(x)		x	
-limitations noted						x	x	
-parallel analysis					x		x	x
-capitalization on chance							x	
7 Effect size variability								
-heterogeneity tested						x	x	
-quality aspects	(x)	x	x			x		x
-treatment variation	x	x	x			x		
-outcome variation	x	x	x	x	x	x	x	x
-subject characteristics	x	x	x	x	x	(x)		x
-design/stat.analysis						x		
-contextual/scope variables	x	x	x		x	x	x	x
V Interpretation								
1 Effect size								
-std.dev/percentile	x	x		(x)		x		
-binomial effect size display								
-Cohen's classification					x			
-behavioral indices								
-expert judgement								
2 Theoretical implications								
-old theory/impressions	x	x	x		x	x	x	x
-new theory/hypotheses					(x)		(x)	(x)
3 Practical implications								
-for policy or practice	x	x	x	x				x
-limitations noted	x	x	x	(x)	x	x	x	x
4 Future implications								
-for primary research	(x)		x	x	x	x	x	x
-for reviews			(x)	x		x		

3.A. Frequency of recordings per year

	1979/80	1981	1982	1983	1984	1985-88	Tot
Number of meta-analyses	7	10	10	12	9	7	55
I Theoretical framework							
1 -problem orientated	4	9	7	10	6	6	42
-indirectly deducible	3	1	3	3	3	2	15
2 -prior reviews mentioned	7	9	8	11	6	6	47
-prior research mentioned	5	6	5	8	6	7	37
-shortcomings indicated	5	4	5	8	4	4	30
3 Construct definitions							
-specifically formulated	4	9	6	9	8	5	41
-indirectly deducible	6	4	7	7	5	3	32
-diversity ignored	1	-	-	1	1	-	3
-diversity analyzed	4	6	8	9	5	5	37
4 Aims							
-exploratory	5	10	9	12	8	6	50
-specific hypotheses	-	1	1	3	3	1	9
II Sampling							
1 Search strategy							
-retrieval system	4	7	7	12	6	4	40
-bibliographies	5	6	6	8	5	6	36
-experts quizzed	-	1	-	1	2	1	5
2 Representativeness							
-published/unpublished	4	4	5	8	6	3	30
-file drawer/fail safe	2	1	1	3	-	2	9
-source variation	3	4	4	6	5	3	25
3 Selection criteria							
-specified	3	10	6	9	5	3	36
-excluded studies	2	5	1	3	4	2	17
-articles listed	1	8	4	8	6	5	32
-available on request	6	1	4	2	2	-	15
III Coded characteristics							
1 sample size	2	4	5	4	3	3	21
2 person characteristics							
-sex	2	4	5	9	2	1	23
-age/class/grade	5	8	9	11	7	7	47
-educational variables	4	4	6	8	3	2	27
-demographic variables	1	4	4	7	3	2	21

3-A continued	1979/80	1981	1982	1983	1984	1985-88	Tot
3 settings							
-laboratory	-	3	1	3	2	4	13
-field	3	8	4	7	4	7	33
-region	2	3	2	4	3	2	16
-publication date	4	9	10	8	8	3	42
4 study features							
-global quality index	-	4	3	3	3	3	16
-sampling/assignment	5	2	5	5	5	4	26
-operationalizations	5	9	10	11	8	6	49
-reliability of outcome	5	7	6	7	3	3	31
-interaction/control var.	3	4	5	4	3	1	20
-unit of analysis	-	2	-	1	1	2	6
-adequacy of stat.anal.	-	3	1	-	4	3	11
IV Data analysis							
1 Effect size conceptualization							
-formula/index indicated	7	10	9	12	8	6	52
-magnitudes listed per study	-	2	3	3	5	5	18
-stem-leaf/graph summary	5	4	6	3	4	2	24
-significance indicated	1	2	2	3	3	1	12
-outcome variable indicated	2	5	6	6	5	6	30
2 Coding quality							
-missing data problem noted	6	8	6	8	5	3	36
-information gained elsewhere	-	1	2	3	1	2	9
-studies excluded	1	5	4	4	3	3	20
-global intercoder agreement	-	5	2	1	2	2	12
-indices per item	-	-	2	-	-	-	2
-strategy to reach agreement	2	1	2	3	-	2	10
3 Unit of analysis							
-study	2	3	4	4	2	4	19
-study findings	5	8	5	9	5	3	35
4 Non-independence							
-present,taken into account	4	5	7	8	3	7	34
-ignored, authors	-	7	4	7	6	5	29
-ignored, findings	5	6	3	7	6	1	28
5 Average effect sizes for							
-total sample	4	4	5	7	6	4	30
-subsample	7	10	8	12	8	6	51
-range/std.dev./std.error	7	8	9	11	6	6	47
6 Statistical analysis							
-conventional	6	9	6	8	6	4	39
-'modern'	-	1	2	7	2	2	14
-assumptions tested	-	-	1	2	1	1	5
-limitations noted	2	3	5	2	2	2	16
-parallel analysis	1	4	3	5	3	1	17
-capitalization on chance	1	1	2	-	1	-	5

3-A continued	1979/80	1981	1982	1983	1984	1985-88	Tot
7 Effect size variability							
-heterogeneity tested	-	1	3	3	2	2	11
-quality aspects	4	7	4	5	4	3	27
-treatment variation	4	10	8	7	7	6	42
-outcome variation	7	8	8	11	7	5	46
-subject characteristics	5	6	8	12	5	4	40
-design/stat.analysis	4	4	3	3	3	3	20
-contextual/scope variables	7	8	8	9	6	4	42
V Interpretation							
1 Effect size							
-std.dev/percentile	7	6	7	8	3	5	36
-binomial effect size display	-	-	-	-	1	-	1
-Cohen's classification	2	2	4	4	2	-	14
-behavioral indices	1	1	-	1	1	1	5
-expert judgement	-	-	-	-	-	1	1
2 Theoretical implications							
-old theory/impressions	5	8	7	11	6	6	43
-new theory/hypotheses	-	1	1	3	-	-	5
3 Practical implications							
-for policy or practice	1	6	1	5	3	3	19
-limitations noted	5	7	7	6	8	5	38
4 Future implications							
-for primary research	4	7	7	9	4	2	33
-for reviews	2	2	3	3	2	2	14

3.B. Percentage of recordings per year

	1979/80	1981	1982	1983	1984	1985-88	Tot
Number of meta-analyses	7	10	10	12	9	7	55
I Theoretical framework							
1 -problem orientated	57	90	70	83	67	86	76
-indirectly deducible	43	10	30	25	33	29	27
2 -prior reviews mentioned	100	90	80	91	67	86	85
-prior research mentioned	71	60	50	67	67	100	67
-shortcomings indicated	71	40	50	67	44	57	55
3 Construct definitions							
-specifically formulated	57	90	60	75	89	71	75
-indirectly deducible	86	40	70	58	56	43	58
-diversity ignored	14	-	-	8	11	-	5
-diversity analyzed	57	60	80	75	56	71	67
4 Aims							
-exploratory	71	100	90	100	89	86	91
-specific hypotheses	-	10	10	25	33	14	16
II Sampling							
1 Search strategy							
-retrieval system	57	70	70	100	67	57	73
-bibliographies	71	60	60	67	56	68	65
-experts quizzed	-	10	-	8	22	14	9
2 Representativeness							
-published/unpublished	57	40	50	67	67	43	55
-file drawer/fail safe	29	10	10	25	-	29	16
-source variation	43	40	40	50	56	43	45
3 Selection criteria							
-specified	43	100	60	75	56	43	65
-excluded studies	29	50	10	25	44	29	31
-articles listed	14	80	40	67	67	71	58
-available on request	86	10	40	17	22	-	27
III Coded characteristics							
1 sample size	29	40	50	33	33	43	38
2 person characteristics							
-sex	29	40	50	75	22	14	42
-age/class/grade	71	80	90	91	78	100	85
-educational variables	57	40	60	67	33	29	49
-demographic variables	14	40	40	58	33	29	38

3-B continued	1979/80	1981	1982	1983	1984	1985-88	Tot
3 settings							
-laboratory	-	30	10	25	22	57	24
-field	43	80	40	58	44	100	60
-region	29	30	20	33	33	29	29
-publication date	57	90	100	67	89	43	76
4 study features							
-global quality index	-	40	30	25	33	43	29
-sampling/assignment	71	20	50	42	56	57	47
-operationalizations	71	90	100	91	89	86	89
-reliability of outcome	71	70	60	58	33	43	56
-interaction/control var.	43	40	50	33	33	14	36
-unit of analysis	-	20	-	8	11	29	11
-adequacy of stat.anal.	-	30	10	-	44	43	20
IV Data analysis							
1 Effect size conceptualization							
-formula/index indicated	100	100	90	100	89	86	95
-magnitudes listed per study	-	20	30	25	56	71	33
-stem-leaf/graph summary	71	40	60	25	44	29	44
-significance indicated	14	20	20	25	33	14	22
-outcome variable indicated	29	50	60	50	56	86	55
2 Coding quality							
-missing data problem noted	86	80	60	67	56	43	65
-information gained elsewhere	-	10	20	25	11	29	16
-studies excluded	14	50	40	33	33	43	36
-global intercoder agreement	-	50	20	8	22	29	22
-indices per item	-	-	20	-	-	-	4
-strategy to reach agreement	29	10	20	25	-	29	18
3 Unit of analysis							
-study	29	30	40	33	22	57	35
-study findings	71	80	50	75	56	43	64
4 Non-independence							
-present,taken into account	57	50	70	67	33	100	62
-ignored, authors	-	70	40	58	67	71	53
-ignored, findings	71	60	30	58	67	14	51
5 Average effect sizes for							
-total sample	57	40	50	58	67	57	55
-subsample	100	100	80	100	89	86	93
-range/std.dev./std.error	100	80	90	91	67	86	85
6 Statistical analysis							
-conventional	86	90	60	67	67	57	71
-'modern'	-	10	20	58	22	29	25
-assumptions tested	-	-	10	17	11	14	9
-limitations noted	29	30	50	17	22	29	29
-parallel analysis	14	40	30	42	33	14	31
-capitalization on chance	14	10	20	-	11	-	9

3-B continued	1979/80	1981	1982	1983	1984	1985-88	Tot
7 Effect size variability							
-heterogeneity tested	-	10	30	25	22	29	20
-quality aspects	57	70	40	42	44	43	49
-treatment variation	57	100	80	58	78	86	76
-outcome variation	100	80	80	91	78	71	84
-subject characteristics	71	60	80	100	56	57	73
-design/stat.analysis	57	40	30	25	33	43	36
-contextual/scope variables	100	80	80	75	67	57	76
V Interpretation							
1 Effect size							
-std.dev/percentile	100	60	70	67	33	71	65
-binomial effect size display	-	-	-	-	11	-	2
-Cohen's classification	29	20	40	33	22	-	25
-behavioral indices	14	10	-	8	11	14	9
-expert judgement	-	-	-	-	-	14	2
2 Theoretical implications							
-old theory/impressions	71	80	70	91	67	86	78
-new theory/hypotheses	-	10	10	25	-	-	9
3 Practical implications							
-for policy or practice	14	60	10	42	33	43	35
-limitations noted	71	70	70	50	89	71	69
4 Future implications							
-for primary research	57	70	70	75	44	29	60
-for reviews	29	20	30	25	22	29	25